



**CENTRO UNIVERSITÁRIO INSTITUTO DE
EDUCAÇÃO SUPERIOR DE BRASÍLIA**

**Bacharelado em
Ciência de Dados e Inteligência Artificial**

Trabalho de Conclusão de Curso - TCC

Victor Augusto Souza Resende

**Aplicações de machine learning para a predição e
controle de incêndios florestais: uma análise a partir
de índices de secas**

**Brasília
2023**

Aplicações de machine learning para a predição e controle de incêndios florestais: uma análise a partir de índices de secas

Trabalho de Conclusão de Curso (TCC) apresentado como pré-requisito para obtenção de Bacharelado em Ciência de Dados e Inteligência Artificial pelo Instituto de Educação Superior de Brasília - IESB.

Orientador: Professor Sérgio da Costa Côrtes

Aplicações de machine learning para a predição e controle de incêndios florestais: uma análise a partir de índices de secas

Victor Augusto Souza Resende

Trabalho de Conclusão de Curso (TCC) apresentado como pré-requisito para obtenção de Bacharelado em Ciência de Dados e Inteligência Artificial pelo Instituto de Educação Superior de Brasília - IESB

Brasília – Distrito Federal, em _____ de _____ de _____

Banca Examinadora constituída pelos Professores:

Prof^a. Dr^a. Fulana de Tal - Orientadora
Professor no IESB

Prof. Dr. Fulano de Tal - Membro
Professor no IESB

Prof. Dr. Fulano de Tal - Membro Externo
Professor na UnB

Prof. Dr. Fulano de Tal - Membro Externo
Professor na ENCE

Agradecimentos

Agradeço a Deus por proporcionar todos os momentos que até então me tornaram quem sou hoje. Aos meus pais que sempre me apoiaram e incentivaram-me em todas as decisões das quais decidi e pela educação que me proporcionaram. Agradeço a minha namorada por toda motivação, apoio e acolhimento na criação da pesquisa. Sou grato por toda perseverança e empenho da qual batalhei durante a graduação através das oportunidades que tive em meu início de carreira. Gratifico todos os meus amigos que participaram da minha rotina, sem eles essa jornada não teria sido tão especial quanto foi. Agradeço, também, a todos os professores do IESB que com seus conhecimentos, competências e habilidades me ensinaram e me prepararam para ser um profissional de excelência e um ser humano melhor.

*“No que diz respeito ao empenho,
ao compromisso, ao esforço,
a dedicação, não existe meio termo.
Ou você faz uma coisa bem feita
ou não faz.”
(Ayrton Senna)*

Resumo

Atualmente a seca é um dos desastres naturais com maior ocorrência no mundo, principalmente após as mudanças climáticas nos últimos anos. Entretanto, existem diversos cálculos para identificar a severidade de determinada seca, como os índices SPI e SPEI. Além disso, tal comportamento climático pode ter relação com a quantidade de queimadas em determinada região, dada as condições ideias para a propagação do fogo. Portanto, o estudo aborda a criação de um modelo estatístico a fim de prever a frequência mensal de incêndios a partir de variáveis climáticas e índices de secas na localidade de Brasília-DF no período de 1992 a 2022.

Palavras-chaves: Ciência de Dados, Inteligência Artificial, *Aprendizado de Máquina*, SPI, SPEI, incêndios florestais, queimadas.

ABSTRACT

Drought is currently one of the most common natural disasters in the world, especially after climate change in recent years. However, there are several calculations to identify the severity of a certain drought, such as the SPI and SPEI indices. In addition, such climatic behaviour may be related to the number of fires in a given region, given the ideal conditions for the spread of fire. Therefore, the study addresses the creation of a statistical model in order to predict the monthly frequency of fires from climatic variables and drought indices in the locality of Brasília-DF in the period from 1992 to 2022.

Keywords: Data Science, Artificial Intelligence, *Machine Learning*, SPI, SPEI, forest fires.

Lista de ilustrações

Figura 1 – Estação convencional INMET - Brasília, DF 83377	24
Figura 2 – Satélite AQUA	24
Figura 3 – Modelo Entidade Relacionamento e Diagrama Entidade Relacionamento	25
Figura 4 – Precipitação e temperaturas - Estatísticas básicas	27
Figura 5 – Fluxo de desenvolvimento	40
Figura 6 – Série histórica - Precipitação	42
Figura 7 – Histograma - Precipitação	42
Figura 8 – Série histórica - Temperatura máxima média	43
Figura 9 – Histograma - Temperatura máxima média	44
Figura 10 – Série histórica - Temperatura mínima média	44
Figura 11 – Histograma - Temperatura mínima média	45
Figura 12 – Série histórica - Temperatura média mensal	46
Figura 13 – Histograma - Temperatura média	46
Figura 14 – Série histórica - Incêndios florestais	47
Figura 15 – Série histórica mensal - Incêndios florestais	48
Figura 16 – Relacionamento Precipitação e Queimadas	48
Figura 17 – Matriz de correlação	49
Figura 18 – Boxplot das variáveis de estudo	51
Figura 19 – Série histórica índice SPI6	52
Figura 20 – Distribuição índices SPI6	53
Figura 21 – Série histórica índice SPEI3	54
Figura 22 – Distribuição índices SPEI3	55
Figura 23 – Matriz de correlação	57
Figura 24 – Introdução	62
Figura 25 – Visualização das séries históricas	62
Figura 26 – Previsões	62

Lista de tabelas

Tabela 1 – Cronograma	22
Tabela 2 – Classificação do Índice de precipitação padronizada (SPI)	35
Tabela 3 – Quantidade de dados nulos por variável	50
Tabela 4 – Comparação dos modelos	61

Listas de abreviaturas e siglas

API	Interface de Programação de Aplicativos (<i>Application Programming Interface</i>)
INMET	Instituto Nacional de Meteorologia
INPE	Instituto Nacional de Pesquisas Espaciais
SPI	Índice de Precipitação Padronizada (<i>Standardized Precipitation Index</i>)
SPEI	Índice Padronizado de Precipitação-Evapotranspiração (<i>Standardised Precipitation Evapotranspiration Index</i>)

Sumário

1	INTRODUÇÃO	19
2	MOTIVAÇÃO	20
3	OBJETIVOS E ORGANIZAÇÃO DO TRABALHO	21
3.1	Objetivos	21
3.2	Organização do Trabalho	21
3.3	Cronograma do Trabalho	22
4	CONSTRUÇÃO DO BANCO DE DADOS	23
4.1	Definição das Fontes de Dados	23
4.2	Descrição dos Dados nas Fontes de Dados	23
4.3	Modelagem dos Dados e Criação do Banco de Dados	24
4.3.1	Modelos de dados	25
4.3.2	Dicionário de Dados	25
4.3.3	Instituto Nacional de Pesquisas Espaciais (INPE) - Incêndios florestais	26
4.3.4	Instituto Nacional de Meteorologia (INMET) - Estações Convencionais	26
4.3.5	Processo de Carga dos Dados	26
4.3.6	Resumo Estatístico dos Dados no Banco de Dados	26
5	REFERENCIAL TEÓRICO	28
5.1	Secas	28
5.1.1	Conceitos meteorológicos	29
5.1.1.1	Precipitação	29
5.1.1.2	Temperatura	29
5.1.1.3	Temperatura mínima	29
5.1.1.4	Temperatura máxima	30
5.1.1.5	Evapotranspiração	30
5.1.2	Conceitos estatísticos	30
5.1.2.1	Medidas de tendência central	30
5.1.2.2	Medidas de dispersão	30
5.1.2.3	Formato e distribuição dos dados	30
5.1.3	Índice de Precipitação Padronizada	31
5.1.4	Índice Padronizado de Precipitação-Evapotranspiração	35
5.2	Modelagem e inferência	38
5.2.1	Régressão linear	38
5.2.2	Random Forest	39

6	DESENVOLVIMENTO DO PROJETO	40
6.1	Fluxo de desenvolvimento	40
6.2	Análise Exploratória dos Dados	41
6.2.1	Precipitação mensal	41
6.2.2	Temperatura máxima média mensal	43
6.2.3	Temperatura mínima média mensal	44
6.2.4	Temperatura média mensal	45
6.2.5	Incêndios florestais	47
6.2.6	Relação precipitação e frequência de incêndios	48
6.2.7	Correlação	49
6.2.8	Qualidade dos dados	50
6.3	SPI	51
6.4	SPEI	53
6.5	Modelagem dos dados	56
6.5.1	Preparação dos dados	56
6.5.2	Seleção das variáveis	57
6.6	Implementação dos modelos	58
6.6.1	Régressão linear	59
6.6.2	Random forest	59
6.6.3	Rede neural	60
6.6.4	Comparação dos modelos	60
6.7	Aplicação do modelo selecionado em uma aplicação web	61
6.8	Resultados	63
7	CONCLUSÕES	65
	REFERÊNCIAS	66
	ANEXOS	69
	ANEXO A – DICIONÁRIO DE DADOS	70
A.1	Instituto Nacional de Pesquisas Espaciais (INPE) - Incêndios florestais	70
A.2	Instituto Nacional de Meteorologia (INMET) - Estações Convencionais	70
	ANEXO B – CÓDIGOS DOS PROGRAMAS	71
B.1	Inserção e acesso dos dados no banco de dados - Python	71
B.2	Criação e análise índices SPI6 e SPEI3 - R	73
B.3	Análise exploratória e modelagem - Python	75

1 INTRODUÇÃO

Fenômenos naturais são eventos que ocorrem sem a intervenção humana. Existem diversos tipos de fenômenos naturais, como: vulcões, terremotos, tsunamis, secas, enchentes, aurora boreal e arco-íris. Entretanto, alguns dos citados anteriormente podem afetar a vida humana, dos quais assim nomeiam-se “desastres naturais” quando há a ocorrência. À vista disso, tais desastres podem causar grandes impactos econômicos e sociais para uma determinada localidade.

Os eventos climáticos, tais como alagamentos, enxurradas, inundações e chuvas intensas, geraram um prejuízo total de cerca de 30 bilhões de reais na economia, além de danos materiais de aproximadamente 31 bilhões de reais. Separadamente, as enxurradas desembolsaram cerca de R\$ 15.578.851.584,23 em danos materiais e R\$ 11.249.841.426,07 em prejuízos totais ([SALVADOR, 2021](#)). Nos últimos 20 anos, houve um aumento significativo do interesse em eventos de incêndios florestais, não apenas pelos impactos econômicos, mas também pelas consequências ambientais que esses eventos provocam ([SOUZA et al., 2012](#)).

Nesse texto, serão avaliados dados meteorológicos no território da capital brasileira, Brasília-DF, cujos registros foram coletados e disponibilizados pelo Instituto Nacional de Meteorologia (INMET) através de estações convencionais no período de 1992 a 2022. Para tal, haverá a avaliação, principalmente, de índices, como o Índice de Precipitação Padronizado (SPI) e o Índice de Precipitação Evapotranspiração Padronizado (SPEI) e características climáticas pertinentes, a fim de identificar a intensidade, ou severidade, da seca nesse território. Além disso, utilizou-se dados da série histórica de incêndios florestais em Brasília-DF no período de 1998 a 2022, disponibilizados pelo Instituto Nacional de Pesquisas Espaciais (INPE) para verificar a relação entre secas e focos de incêndio na respectiva localidade.

Portanto, outro ponto de destaque do estudo se desenvolve sobre como tais índices voltados a secas podem auxiliar na prevenção e previsão de incêndios florestais mediante modelagem estatística. Além disso, pretende-se realizar a relação dos índices de seca com as frequências dos focos de incêndio na série histórica, considerando outros estudos, dos quais serão citados ao longo do desenvolvimento. Por fim, a pesquisa a seguir visa investigar os fenômenos das secas e estiagens, e como técnicas de aprendizado de máquina, estatística, geografia e matemática aplicada podem auxiliar na prevenção e predição de respectivos focos de incêndios.

2 MOTIVAÇÃO

Com o passar dos anos e o avanço da tecnologia, houve a criação e implementação de diversos sensores e sistemas em locais que correm riscos de desastres naturais para a verificação do perigo eminent e coleta de tais dados. Dessa forma, uma vasta quantidade de dados é gerada diariamente, principalmente quando há situações de estiagem, secas ou enchentes, dos quais são úteis para planos econômicos e sociais. Consequentemente, diversas técnicas de tratamento dos dados, modelagem e inferência estatística podem auxiliar na prevenção do impacto de determinado desastre natural, caso ocorra.

As mudanças climáticas resultantes do aquecimento global têm causado efeitos devastadores na natureza, especialmente no clima. Uma série de variações significativas a longo prazo são observadas em escalas continental, regional e oceânica, incluindo alterações na temperatura e no gelo do Ártico, quantidade de precipitação, salinidade oceânica, padrões de vento e em aspectos de eventos climáticos extremos, tais como secas, chuvas intensas, ondas de calor e intensidade de ciclones tropicais ([JURAS, 2008](#)).

Há uma crescente quantidade de estudos que apontam as emissões de dióxido de carbono e outros gases como responsáveis pelo aumento de temperatura observado recentemente. Ainda não há certeza sobre como a biosfera irá responder à acumulação desses gases, mas grande parte da comunidade científica internacional concorda que é provável ocorrer um aumento no nível dos oceanos. De fato, é considerado altamente provável que ocorra um aumento médio de temperatura global entre 1ºC e 5ºC nos próximos 50 anos ([CHANG; HUNSAKER; DRAVES, 1992](#)). Consequentemente, a seca ocasionada pelas mudanças climáticas é um fator que pode contribuir significativamente para o aumento dos focos de incêndio. Dessa forma, é fundamental que sejam adotadas medidas preventivas, estudos e estratégias de monitoramento, a fim de minimizar os efeitos da seca e evitar a ocorrência de incêndios florestais.

Portanto, o texto a seguir tem como motivação a demonstração de técnicas, métricas e índices estatísticos e meteorológicos para o auxílio e prevenção da intensidade de desastres naturais referentes às secas, como os índices SPI e SPEI, do qual a aplicação e análise realizada sobre tais valores possuem potencial de monitoramento e prevenção de incêndios, dos quais afetam cada vez mais a sociedade, aliado ao relacionamento com políticas públicas de combate às secas, articulados em projetos de leis.

3 Objetivos e Organização do Trabalho

O capítulo a seguir visa apresentar os objetivos da pesquisa, bem como a organização e cronograma de finalização dos resultados obtidos neste Trabalho de Conclusão de Curso (TCC) a fim de auxiliar o leitor sobre a apreciação do desenvolvimento da pesquisa realizada sobre desastres naturais.

3.1 Objetivos

O texto possui como objetivos a demonstração e implementação de técnicas estatísticas, matemáticas e computacionais das quais podem ser úteis na predição e prevenção de desastres naturais relacionados a estiagem e focos de incêndio. Aliado a isso, será apresentado a importância de tais análises por meio de estudos relacionados, visando demonstrar como tais índices podem ajudar na prevenção e monitoramento de focos de incêndio na localidade de Brasília-DF. Por fim, decidiu-se por uma simples implementação web a fim de tornar os resultados interativos com os cidadãos.

3.2 Organização do Trabalho

O trabalho a seguir será organizado de forma que seja explicado trabalhos relacionados, conceitos geográficos sobre fenômenos e desastres naturais, conceitos e métricas estatísticas utilizadas, ferramentas computacionais, desenvolvimento e a apresentação dos resultados e conclusão. Além disso, serão apresentados o desenvolvimento da tese, bem como os respectivos resultados e conclusões.

1. **Introdução:** Apresentação do estudo proposto e desafios a serem sanados.
2. **Motivação:** Razões das quais torna a pesquisa desenvolvida benéfica à sociedade.
3. **Objetivos e organização do trabalho:** Demonstração dos objetivos dos quais busca-se alcançar, organização e cronograma do trabalho.
4. **Construção do Banco de dados:** Construção do ambiente dos quais os dados serão inseridos e consultados para as análises.
5. **Referencial teórico:** Etapa da qual apresentará conceitos dos quais serão utilizados para a elaboração do texto.
 - a) Conceitos meteorológicos sobre temperatura e pluviosidade;
 - b) Conceitos matemáticos e estatístico referentes aos índice SPI e SPEI;

- c) Conceitos e métricas estatísticas;
 - d) Conceitos computacionais sobre aprendizado de máquina.
6. **Desenvolvimento:** Etapa da qual apresentará a pesquisa sobre as bases de dados referente aos desastres naturais citados anteriormente.
- a) Análise exploratória;
 - b) Manipulação das bases de dados;
 - c) Criação de modelos para a predição dos incêndios florestais a partir das variáveis climáticas.
7. **Resultados:** Apresentação dos resultados encontrados após a pesquisa aplicada sobre as bases de dados propostas.
8. **Conclusões:** Encerramento do trabalho apresentando conclusões das quais podem ser alcançadas através do estudo desenvolvido.

3.3 Cronograma do Trabalho

A especificação do cronograma de atividades visa indicar quais serão os ciclos da pesquisa e quantificar o tempo para finalização de cada uma delas. Para isto, é necessário um resumo de todas as etapas que compõem o texto. Portanto, o cronograma para a confecção desta pesquisa pode ser visto abaixo.

Tabela 1 – Cronograma

Cronograma pesquisa	Setembro	Outubro	Novembro	Dezembro	Janeiro	Fevereiro	Março	Abril	
Introdução	X								
Motivação	X								
Objetivos		X							
Construção banco de dados		X	X						
Referencial teórico				X	X				
Desenvolvimento					X	X	X		
Resultados								X	X
Conclusões									X

Fonte: autor, 2023

Vale ressaltar que o tempo necessário para finalizar cada ciclo do cronograma de atividades depende dos recursos materiais e humanos do qual o autor possui. Além disso, é válido frisar que tais apontamentos são referentes ao segundo semestre de 2022 e primeiro semestre de 2023.

4 Construção do Banco de Dados

A etapa de construção do banco de dados visa a criação de todas as entidades e relacionamentos que serão necessários para a consulta, disponibilização e análise dos dados realizadas no texto. Ressalta-se que o banco de dados utilizado para o encapsulamento dos dados é referente a uma instância PostgreSQL hospedada no ambiente de nuvem na AWS, e a criação das tabelas foi efetuada através da linguagem de programação Python e suas respectivas bibliotecas.

4.1 Definição das Fontes de Dados

Como citado anteriormente, os dados do artigo são referentes a análises de secas e incêndios florestais que serão realizadas nos próximos capítulos. Portanto, os registros utilizados para a realização desses estudos estão relacionados a pluviosidade, evapotranspiração, precipitação e temperatura na cidade de Brasília-DF no período mensal de 30 anos (1992 a 2022). Da mesma forma, coletou-se dados sobre incêndios florestais no período mensal de 24 anos (1998 a 2022).

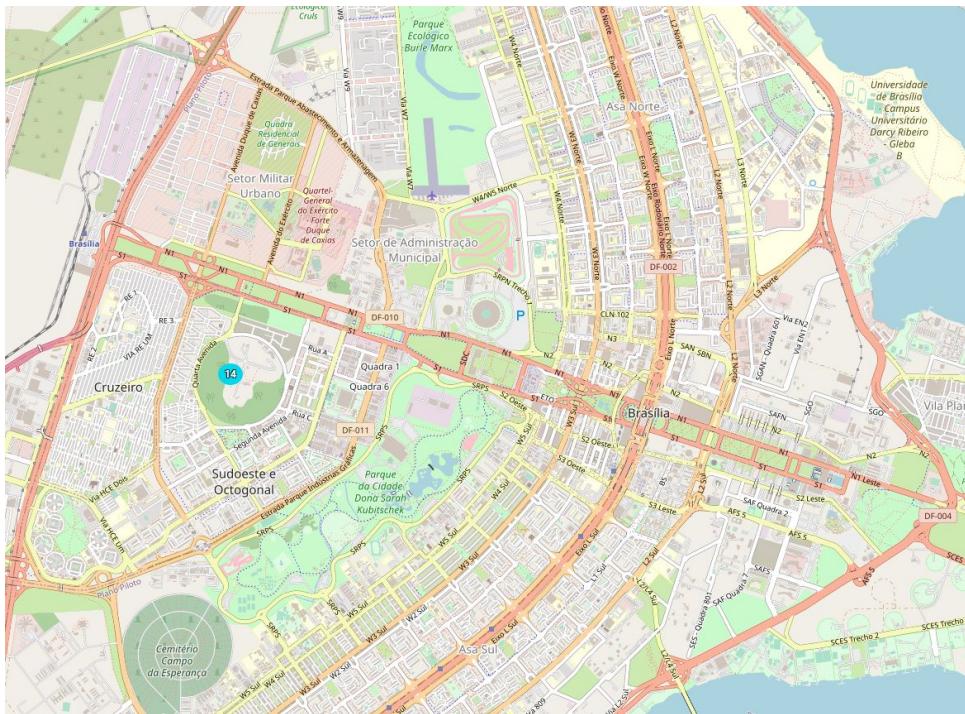
Ademais, as fontes de dados sobre pluviosidade, evapotranspiração, precipitação e temperatura foram coletados através do Instituto Nacional de Meteorologia (INMET) da qual disponibiliza esses por meio de um formulário que o usuário deve preencher e pode ser acessado [clicando aqui](#), referenciando o tipo de estação, localidade, intervalo e granularidade de tempo e as variáveis requisitadas.

Por último, os registros referentes a incêndios florestais foram coletados por meio de requisição através do Instituto Nacional de Pesquisas Espaciais (INPE) e pode ser acessado [clicando aqui](#), dos quais o instituto efetua imagens de satélites para a contabilização dos focos.

4.2 Descrição dos Dados nas Fontes de Dados

Essa seção visa demonstrar as variáveis contidas nas fontes de dados utilizadas no texto, das quais estão disponíveis no dicionário de dados. Vale ressaltar que a estação do INMET escolhida foi a estação convencional, localizada na cidade de Brasília-DF, da qual possui a numeração 83377, onde, na figura abaixo, é possível identificá-la por meio do ponto azul.

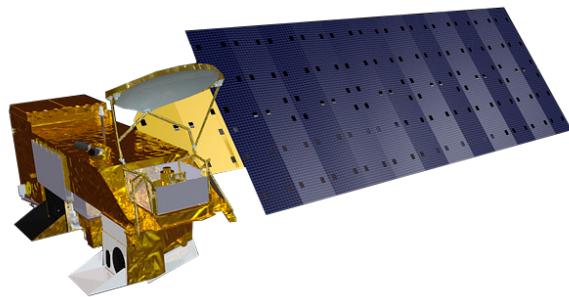
Figura 1 – Estação convencional INMET - Brasília, DF 83377



Fonte: Google Maps, 2023

Os dados referentes a análise de focos de incêndios florestais requisitados pelo INPE são contabilizados por meio de imagens do satélite denominado “AQUA”, criado em uma parceria entre as nações do Brasil e Japão, abaixo é possível ver uma imagem meramente ilustrativa.

Figura 2 – Satélite AQUA



Fonte: Instituto Nacional de Pesquisas Espaciais, 2023

4.3 Modelagem dos Dados e Criação do Banco de Dados

A modelagem de dados é o processo da criação de uma estrutura de informações que serão capturadas por um banco de dados. Portanto, após a criação do conceito que será utilizado, é realizado a implementação do banco de dados, do qual deverá seguir

regras e princípios estabelecidos na modelagem, referenciando os dados e seus respectivos relacionamentos.

4.3.1 Modelos de dados

De maneira objetiva, os modelos de dados são representações visuais dos elementos de dados, seus atributos e respectivos relacionamentos com outras entidades. Entretanto, as bases de dados utilizadas nesse artigo não possuem relacionamento entre si, isto é, a tabela temperaturas não possui relacionamento com a tabela incêndios, uma vez que foram propostas análises apartadas sobre os dados de cada uma, relacionando apenas o impacto. Dessa forma, abaixo é representado o Modelo Entidade Relacionamento (MER) e Diagrama de Entidade e Relacionamento (DER) das entidades das quais serão utilizadas na elaboração da ideia proposta.

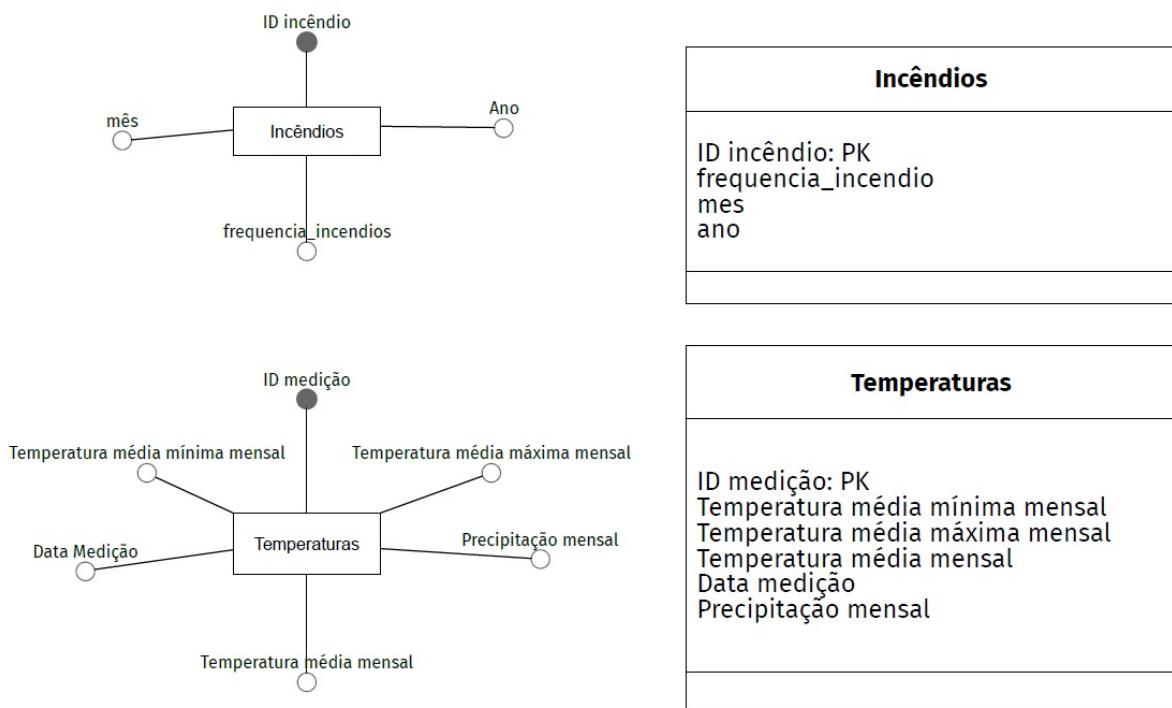


Figura 3 – Modelo Entidade Relacionamento e Diagrama Entidade Relacionamento

4.3.2 Dicionário de Dados

O dicionário de dados visa explicar as variáveis presentes no banco de dados abordado neste projeto. Dessa forma, como citado anteriormente, o texto considera o conjuntos de dados, referente às variáveis sobre pluviosidade e temperatura. Portanto, a seguir é apresentado os respectivos atributos desse conjunto de dados.

4.3.3 Instituto Nacional de Pesquisas Espaciais (INPE) - Incêndios florestais

- **Ano:** Ano da respectiva contabilização mensal dos incêndios.
- **Mês:** Mês da respectiva contabilização mensal dos incêndios.
- **Frequência de incêndios:** A respectiva contabilização mensal dos incêndios.

4.3.4 Instituto Nacional de Meteorologia (INMET) - Estações Convencionais

- **Data Medição:** Data referente a coleta dos dados. Formato (YYYY-MM-DD).
- **Precipitação total (diário):** Total de precipitação (chuva) que ocorreu no espaço de 24 hora. Medida em milímetros (mm).
- **Temperatura Mínima (diário):** Mínima da temperatura do ar, ocorrida no período de 24 horas, a partir do termômetro específico para este fim. Medida em graus Celsius.
- **Temperatura Máxima (diário):** Máxima da temperatura do ar, ocorrida no período de 24 horas, a partir do termômetro específico para este fim. Medida em graus Celsius.
- **Temperatura Média (diário):** Média da temperatura do ar, ocorrida no período de 24 horas, a partir do termômetro específico para este fim. Medida em graus Celsius.

4.3.5 Processo de Carga dos Dados

De maneira inicial, o processo de carga dos dados no banco de dados se desenvolveu escolhendo o intervalo de tempo do qual o estudo abordará. Dessa forma, foi-se coletado 30 anos dos registros referentes a pluviosidade e temperatura no intervalo de 1992 a 2022. Com relação aos registros sobre incêndios florestais utilizou-se 24 anos, de 1998 a 2022.

Consequentemente, em relação aos dados de pluviosidade e temperaturas criou-se a tabela **temperaturas**, já para incêndios florestais a tabela **incêndios**. Portanto, a inserção dos dados no banco de dados foi realizada através da biblioteca SQLAlchemy presente na linguagem de programação Python.

4.3.6 Resumo Estatístico dos Dados no Banco de Dados

A fim de verificar os dados extraídos e inseridos nos bancos de dados, decidiu-se pela criação de um breve resumo estatístico de tais registros, principalmente visando a

verificação da qualidade de tais dados. Vale ressaltar que na etapa de desenvolvimento haverá uma análise exploratória mais ampla diante dos dados coletados.

Em relação aos dados de precipitação e temperaturas, identificou-se a inserção de 370 registros mensais no período de 30 anos de coleta de Dados, considerando o período de 1992 a 2022 referentes ao território de Brasília-DF. Abaixo é possível identificar as estatísticas básicas sobre tais dados quantitativos.

	PRECIPITACAO	TEMPERATURA_MAXIMA_MEDIA	TEMPERATURA_MEDIA	TEMPERATURA_MINIMA_MEDIA
count	372.000000	372.000000	372.000000	372.000000
mean	122.013978	27.105747	21.408394	16.993051
std	118.072568	1.578804	1.397325	1.771593
min	0.000000	23.383871	17.681290	11.903226
25%	9.850000	26.006452	20.535968	15.420968
50%	92.400000	26.877043	21.498988	17.763334
75%	204.150000	27.943199	22.263467	18.322769
max	526.400000	33.000000	25.623226	19.729032

Figura 4 – Precipitação e temperaturas - Estatísticas básicas

De maneira inicial, pode-se verificar que a variável referente a precipitação no período dos 30 anos de coleta de dados apresentou média mensal de 122.01 milímetros (mm), com a precipitação máxima mensal alcançando 526.4 milímetros (mm) no mês de outubro de 2006. Além disso, pode-se afirmar que pelo menos metade dos meses analisados no período alcançou uma precipitação de 92.4 milímetros (mm). Por fim, o desvio padrão mensal da precipitação, em milímetros, foi de 118.07.

Em relação aos dados de temperatura, vale ressaltar que Brasília-DF possui um clima de caráter seco, do qual favorece a amplitude térmica, uma vez que há a falta de umidade para segurar a temperatura, fenômeno do qual é possível identificar na figura 4. Portanto, é possível verificar que a temperatura mensal máxima registrada foi de 33 graus Celsius, já a temperatura mensal média mínima registrada igualou-se a 11 graus Celsius. Por fim, considerando apenas a temperatura média mensal, entende-se que pelo menos metade dos meses coletados, isto é, 50% deles possuem temperatura média menor ou igual a 21.49 graus Celsius.

Por fim, não foi identificado dados menais faltantes para nenhuma das variáveis das quais serão utilizadas no estudo. Dessa forma, conclui-se a existência de 30 anos de registros, do qual se torna necessário para a aplicação dos índices SPI e SPEI, que serão explicados e aplicados posteriormente no desenvolvimento do texto.

5 Referencial Teórico

Em relação aos fenômenos naturais referentes às secas, esses são estudados desde a antiguidade, principalmente visando uma agricultura mais produtiva, de modo que essa era a principal forma de alimentação da população mundial, uma vez que tal situação afeta todos os continentes do planeta. Entretanto, desde as relevantes revoluções industriais, das quais possivelmente acelerou o processo de aquecimento global, a temperatura e a meteorologia do planeta estão sendo afetadas a cada ano, havendo mudanças climáticas catastróficas em uma frequência crescente. Portanto, o estudo sobre a umidade presente no solo faz-se valer a pena visando a utilização positiva do solo, principalmente para áreas como agricultura, moradias e propensas a incêndios.

5.1 Secas

A seca é um fenômeno climático que se diferencia claramente de outras catástrofes naturais. A principal diferença refere-se ao fato de que, diferentemente de outros desastres naturais, como cheias, furacões e terremotos, que geralmente têm um início e fim repentinos e são normalmente restritos a uma pequena região, o fenômeno da seca tem um início lento, longa duração e geralmente se espalha por uma extensa área. ([MOLINA; LIMA, 1999](#)).

Da mesma maneira, diversos autores como diversos autores definem a existência de quatro tipos de secas predominantes, dependendo da temática e abordagem analisada. Logo, a literatura cita que os quatro tipos de secas são referentes às categorias: meteorológica, agrícola, hidrológica e socioeconômica ([MCKEE et al., 1993](#)).

Há diversos estudos e métricas considerados para se compreender a severidade de uma seca em determinada região. Entre esses estudos, a definição de índices quantitativos é a abordagem mais amplamente difundida. No entanto, devido à subjetividade na definição de seca, tem sido extremamente difícil estabelecer um único e universal índice de seca ([VICENTE-SERRANO et al., 2010](#)). Para investigar a seca e seus impactos, incluindo o início e o fim, a abordagem mais eficaz é o uso de índices criados por diferentes pesquisadores para medir a severidade da seca ([FERNANDES et al., 2009](#)).

Entretanto, devido à natureza dos dados coletados e considerando a implementação por meio de técnicas computacionais, o estudo abordará apenas os índices SPI e SPEI. Além disso, os dados coletados e disponibilizados pelo INMET encaixam-se no tipo de seca meteorológica, do qual o texto discorrerá mais detalhadamente sobre.

5.1.1 Conceitos meteorológicos

A meteorologia é o estudo científico do tempo e clima da atmosfera terrestre. Compreender os padrões meteorológicos é crucial para prever as condições climáticas futuras e seus impactos globais. Da mesma forma, a meteorologia tem impacto direto em áreas como agricultura, transportes e turismo, e é fundamental para o gerenciamento de riscos relacionados a eventos climáticos extremos, como furacões, inundações e secas. Além disso, com a crescente preocupação com as mudanças climáticas, a meteorologia tornou-se ainda mais importante para auxiliar, prever e mitigar seus efeitos sobre a população e o meio ambiente. Portanto, a seguir são apresentadas os conceitos meteorológicos por trás das variáveis das quais serão utilizadas para a composição do texto.

5.1.1.1 Precipitação

Segundo o Instituto Nacional de Meteorologia (INMET), o processo de precipitação se desenvolve a partir da ação dos raios solares e do vento sobre as águas da superfície terrestre, provocando o fenômeno da evaporação, que é a passagem da água do estado líquido para o estado de vapor. Devido à evaporação, uma quantidade enorme de gotículas de água fica em suspensão na atmosfera. Gotículas de água se concentram, formando nuvens. Ao se resfriar, a água das nuvens se precipita em forma de chuva. Por este motivo, a chuva é um tipo de precipitação pluvial. A quantidade de chuva num determinado lugar e num determinado tempo, é medida pelo pluviômetro e registrada pelo pluviógrafo. Considera-se precipitação todas as formas de água, líquida ou sólida, que caem das nuvens alcançando o solo: garoa, garoa gelada, chuva fria, granizo, cristais de gelo, bolas de gelo, chuva, neve, bolas de neve e partículas de neve. Seu volume é expresso geralmente em polegadas, referindo-se ao estado da água - se líquida ou sólida - que cai sobre uma determinada região e por um determinado período.

5.1.1.2 Temperatura

Também, de acordo com o Instituto Nacional de Meteorologia (INMET), o conceito de temperatura refere-se a quantidade de calor que existe no ar. Ela é medida pelo termômetro meteorológico, que difere do termômetro clínico. A diferença entre a maior e a menor temperatura chama-se amplitude térmica.

5.1.1.3 Temperatura mínima

Além disso, o Instituto Nacional de Meteorologia (INMET) considera a temperatura mínima nas estações convencionais é a medida mínima da temperatura do ar, ocorrida no período de 24 horas, a partir do termômetro específico para este fim.

5.1.1.4 Temperatura máxima

Também, o Instituto Nacional de Meteorologia (INMET) considera a temperatura máxima nas estações convencionais é a medida máxima da temperatura do ar, ocorrida no período de 24 horas, a partir do termômetro específico para este fim.

5.1.1.5 Evapotranspiração

Por fim, o Instituto Nacional de Meteorologia (INMET) descreve a evapotranspiração como o total de água transferida da superfície da Terra para a atmosfera. É composto da evaporação do líquido, ou “água sólida”, acrescida da transpiração das plantas.

5.1.2 Conceitos estatísticos

A principal vantagem de utilizar estatísticas em estudos de caso se desenvolve por meio do raciocínio analítico que enfoca a onipresença da variação, a investigação das fontes de variação, o planejamento de coleta de dados com a variação em mente, a quantificação da variação e a explicação da variação.

Portanto, a estatística contribui para que dados gerem conhecimento e, deve ter como objetivo não só a produção de dados, como também a interpretação de dados já existentes, utilizando a combinação de gráficos, tabelas e medidas numéricas que permitam interpretar o que esses dados significam.

5.1.2.1 Medidas de tendência central

São medidas que tentam descrever o meio, ou a parte central, de um conjunto de dados. Algumas dessas medidas podem ser catalogadas como: média, mediana e moda.

5.1.2.2 Medidas de dispersão

As medidas de dispersão são medidas que representam a variação dos dados numa base de dados, algumas dessas medidas são catalogadas como: variância e desvio padrão.

5.1.2.3 Formato e distribuição dos dados

O formato da distribuição dos dados tem como dever descrever como os dados são distribuídos, explicando então se a distribuição é simétrica ou assimétrica. Existindo três tipos elementares de distribuição de dados para uma variável. Ao analisar as medidas de tendência central, poderemos dizer em qual tipo de formato citado a variável analisada irá se identificar.

5.1.3 Índice de Precipitação Padronizada

O Índice de Precipitação Padronizada (SPI) é um índice amplamente utilizado para caracterizar a seca meteorológica em uma variedade de escalas de tempo. Em escalas de tempo curtas, o SPI está intimamente relacionado à umidade do solo, enquanto em escalas de tempo mais longas, o SPI pode estar relacionado a águas subterrâneas e armazenamento em reservatórios. O SPI pode ser comparado entre regiões com climas marcadamente diferentes. Ele quantifica a precipitação observada como um desvio padronizado de uma função de distribuição de probabilidade selecionada que modela os dados brutos de precipitação. Os dados brutos de precipitação são tipicamente ajustados a uma distribuição Gama ou de Pearson Tipo III e, em seguida, transformados em uma distribuição normal. Os valores de SPI podem ser interpretados como o número de desvios padrão pelos quais a anomalia observada se desvia da média de longo prazo. Preocupações foram levantadas sobre a utilidade do SPI como uma medida de mudanças na seca associada à mudança climática, uma vez que não lida com mudanças na evapotranspiração. Índices alternativos que tratam da evapotranspiração têm sido propostos, como o Índice de Evapotranspiração e Precipitação Padronizada (SPEI) ([KEYANTASH, 2023](#)).

Além disso, o SPI é calculado para diferentes escalas de tempo, significando o período durante o qual se acumula o valor de precipitação: o SPI1 corresponde à precipitação mensal, o SPI3 corresponde à precipitação acumulada em períodos de 3 meses e assim por diante. SPI - Índice de precipitação padronizada.

Esse índice quantifica o déficit de precipitação para múltiplas escalas de tempo que refletem o impacto da seca na disponibilidade de fontes de água. As condições de umidade de solo respondem às anomalias de precipitação em uma escala de tempo relativamente curta. O armazenamento de água subterrânea, dos fluxos de rios e do reservatório refletem as anomalias de precipitação a longo prazo. Com isso, decidiu-se calcular os índices em escalas de tempo usuais, como: três, seis, doze, vinte e quatro, e quarenta e oito meses ([FERNANDES et al., 2009](#)).

Para o cálculo do SPI, deve-se utilizar uma base de dados de precipitação de ao menos 30 anos, do qual se ajusta mediante uma distribuição gama, que posteriormente é transformada em uma distribuição normal, que por definição, apresenta sua média com valor zero e variância unitária ([FERNANDES et al., 2009](#)). Então, a distribuição gama é determinada pela função de densidade de probabilidade apresentada pela equação 1:

$$g(x) = \frac{1}{\Gamma(\alpha)\beta^\alpha} \chi^{\alpha-1} e^{-\frac{x}{\beta}} \quad (5.1)$$

Sendo:

- $\alpha > 0$: O parâmetro de forma (adimensional);

- $\beta > 0$: O parâmetro de escala (mm);
- $\chi > 0$: O parâmetro de precipitação (mm);
- $\Gamma(\chi)$: A função gama.

Da mesma forma, a função gama é obtida a partir da equação 2:

$$\Gamma(\alpha) = \int_0^{\infty} y^{\alpha-1} e^{-y} dy \quad (5.2)$$

Consequentemente, para estimar os parâmetros α e β apresentados na equação 1, faz-se o uso do método da máxima verossimilhança ([FERNANDES et al., 2009](#)), apresentado pelas equações 3, 4 e 5:

$$A = \ln(\bar{X}) - \frac{1}{N} \sum_{i=1}^N \ln(X) \quad (5.3)$$

$$\alpha = \frac{1}{4A} \left(1 + \sqrt{1 + \frac{4A}{3}} \right) \quad (5.4)$$

$$\beta = \frac{\bar{X}}{\alpha} \quad (5.5)$$

Sendo:

- \bar{X} : média aritmética da precipitação puvial (mm);
- \ln : logaritmo neperiano;
- N : número de observações de precipitação.

Logo, a equação 6 a seguir apresenta a probabilidade cumulativa de um evento de precipitação observado em uma escala de tempo mensal, utilizando os resultados dos parâmetros de forma e escala ([FERNANDES et al., 2009](#)):

$$G(x) = \int_0^x g(x)dx = \frac{1}{\beta^\alpha \Gamma(\alpha)} \int_0^x \chi^{\alpha-1} e^{-\frac{x}{\beta}} dx \quad (5.6)$$

Consequentemente, ao substituir $t = \frac{x}{\beta}$, a equação anterior transforma-se na função gama incompleta ([FERNANDES et al., 2009](#)), resultando na equação 7:

$$G(x) = \frac{1}{\Gamma(\alpha)} \int_0^x t^{\alpha-1} e^{-t} dt \quad (5.7)$$

Ao considerar que a função gama é indeterminada para $x = 0$ e uma distribuição de precipitação pode conter zeros (FERNANDES et al., 2009), a probabilidade cumulativa apresenta o seguinte aspecto apresentado abaixo pela equação 8:

$$H(x) = q + (1 - q)G(x) \quad (5.8)$$

Sendo:

- $H(x)$: distribuição de probabilidade cumulativa;
- q : probabilidade de ocorrência de valores nulos (zeros);
- $G(x)$: distribuição cumulativa teórica.

A transformação equiprobabilística é uma etapa importante no processo de conversão da distribuição gama da probabilidade cumulativa $H(x)$, apresentada na equação 8, para uma distribuição normalizada (Z) com média zero e desvio padrão 1, que corresponderá ao valor de SPI. Essa transformação é essencial para garantir que a probabilidade de ser menor que um valor dado seja igual à probabilidade de ser menor que o valor correspondente da variável transformada. Além disso, os dados de precipitação são ordenados em ordem crescente de magnitude (FERNANDES et al., 2009), e o tamanho da amostra é determinado pela equação 9:

$$q = \frac{m}{n + 1} \quad (5.9)$$

Sendo:

- m : número de ordem dos valores de zero em uma série climatológica;
- n : tamanho da amostra.

Então, o valor de (Z) ou SPI é obtido facilmente pela aproximação matemática que converte a probabilidade cumulativa em uma distribuição normal a variável (Z) (FERNANDES et al., 2009), do qual é demonstrado abaixo pelas equações 10 e 11:

$$Z = SPI = -(t - \frac{C_0 + C_1 t + C_2 t^2}{1 + d_1 t + d_2 t^2 + d_3 t^3}) \text{ para } 0 < H(x) \leq 0.5 \quad (5.10)$$

$$Z = SPI = +\left(t - \frac{C_0 + C_1t + C_2t^2}{1 + d_1t + d_2t^2 + d_3t^3}\right) \text{ para } 0.5 < H(x) \leq 1 \quad (5.11)$$

Sendo t definido pelas equações 13 e 14:

$$t = \sqrt{\ln\left[\frac{1}{(H(x))^2}\right]} \text{ para } 0 < H(x) \leq 0.5 \quad (5.12)$$

$$t = \sqrt{\ln\left[\frac{1}{1 - (H(x))^2}\right]} \text{ para } 0.5 < H(x) \leq 1 \quad (5.13)$$

Além disso, os coeficientes utilizados nas equações 11 e 12 são:

- C_0 : 2.515517;
- C_1 : 0.802853;
- C_2 : 0.010328;
- d_1 : 1.432788;
- d_2 : 0.189269;
- d_3 : 0.001308.

Em termos conceituais, o SPI é uma medida de desvio padrão conhecida como z-score, que indica quantos desvios padrão um evento (ou valor) está acima ou abaixo da média. No entanto, essa interpretação não é completamente precisa para escalas de tempo curtas, já que a distribuição original de precipitação pode não ser simétrica ([FERNANDES et al., 2009](#)).

[McKee et al. \(1993 apud FERNANDES et al., 2009\)](#) utilizaram os valores do SPI para determinar a presença e a intensidade da seca. De acordo com eles, a seca ocorre quando o valor contínuo do SPI é negativo e atinge uma intensidade igual ou inferior a -1,0. O fim da seca é indicado pelo valor positivo do SPI. Essa classificação é realizada utilizando os limites apresentados na Tabela 2, permitindo a caracterização não apenas de secas, mas também de períodos mais úmidos. Essa abordagem tem a grande vantagem de padronizar a análise, permitindo a comparação de regiões completamente diferentes, como áreas com clima mais úmido e chuvoso em relação a áreas mais áridas e secas. ([FERNANDES et al., 2009](#)).

Tabela 2 – Classificação do Índice de precipitação padronizada (SPI)

SPI/SPEI	Classificação
≥ 2.00	Extremamente úmido
1.0 a 1.9	Muito úmido
0.5 a 0.99	Moderadamente úmido
0.49 a -0.49	Próximo ao normal
-0.50 a -0.99	Moderadamente seco
-1.00 a -1.99	Muito seco
≤ -2.00	Extremamente seco

Fonte: Thomas B. McKee et al, 1993

5.1.4 Índice Padronizado de Precipitação-Evapotranspiração

O SPEI cumpre os requisitos de um índice de seca, pois seu caráter multiescalar permite que seja utilizado por diferentes disciplinas científicas para detectar, monitorar e analisar secas. Assim como o sc-PDSI e o SPI, o SPEI pode medir a severidade da seca de acordo com sua intensidade e duração, podendo identificar o início e o fim dos episódios de seca. O SPEI permite a comparação da severidade da seca no tempo e no espaço, pois pode ser calculado em uma ampla gama de climas, assim como o SPI. Além disso, o SPEI atende aos requisitos estabelecidos por (KEYANTASH; DRACUP, 2002) de ser estatisticamente robusto, facilmente calculado e ter um procedimento de cálculo claro e comprehensível. Além disso, uma vantagem crucial do SPEI em relação a outros índices de seca amplamente utilizados é que leva em consideração o efeito do Potencial de Evapotranspiração (PET) na severidade da seca e suas características multiescalares permitem a identificação de diferentes tipos de seca e impactos no contexto do aquecimento global.

Dessa forma, o cálculo do SPEI considera a diferença mensal entre a precipitação e o Potencial de Evapotranspiração (PET). Com um valor para o PET, a diferença entre a precipitação (P) e PET para o mês "i" em questão é calculada (VICENTE-SERRANO et al., 2010), como demonstra a equação 14:

$$D_i = P_i - PET_i \quad (5.14)$$

Dada a similaridade entre as três distribuições (Pearson III, Lognormal e Log-logística) deve-se avaliar a escolha da distribuição estatística mais adequada para modelar a série D. Como apresentado por (VICENTE-SERRANO et al., 2010), a distribuição log-logística mostrou um decréscimo gradual da curva para valores baixos, e probabilidades coerentes foram obtidas para valores muito baixos de D, correspondendo a 1 ocorrência em 200 a 500 anos. Além disso, não foram encontrados valores abaixo do parâmetro origem da distribuição. Dessa forma, a equação 15 demonstra a função de densidade de probabilidade

de uma variável distribuída Log-logística de três parâmetros, como é possível observar na equação 15:

$$f(x) = \frac{\beta}{\alpha} \left(\frac{x - \gamma}{\alpha} \right)^{\beta-1} \left(1 + \left(\frac{x - \gamma}{\alpha} \right)^{\beta} \right)^{-2} \quad (5.15)$$

Onde:

- α : Refere-se ao parâmetro de escala;
- β : Refere-se ao parâmetro de forma;
- γ : Refere-se ao parâmetro de origem.

Os parâmetros da distribuição Log-logística podem ser obtidos seguindo diferentes procedimentos. Entre eles, Ahmad, Sinclair e Werritty (1988 apud VICENTE-SERRANO et al., 2010) demonstram que o procedimento do momento L é a abordagem mais robusta e parcimoniosa. Quando os momentos L são calculados, Upadhyaya e Singh (1999 apud VICENTE-SERRANO et al., 2010) esclarecem que os parâmetros da distribuição Log-logística podem ser obtidos como demonstrado as equações 16, 17 e 18 abaixo:

$$\beta = \frac{2W_1 - W_0}{6W_1 - W_0 - 6W_2} \quad (5.16)$$

$$\alpha = \frac{(W_0 - 2W_1)\beta}{\Gamma(1 + \frac{1}{\beta})\Gamma(1 - \frac{1}{\beta})} \quad (5.17)$$

$$\gamma = W_0 - \alpha\Gamma(1 + \frac{1}{\beta})\Gamma(1 - \frac{1}{\beta}) \quad (5.18)$$

Quando os parâmetros log-logísticos da distribuição α , β , γ foram calculados, foi utilizado o método dos momentos ponderados pela probabilidade (PWMS) obtidos por meio do estimador não enviesado dado por Hosking, Wallis e Wood (1985 apud VICENTE-SERRANO et al., 2010), o desvio padrão da série não muda entre as diferentes escalas de tempo do SPEI. Então, o PWMS não enviesado são obtidos conforme a equação 19:

$$w_s = \frac{1}{N} \sum_N^{i=1} \frac{\left(\frac{N-i}{S}\right) D_i}{\frac{N-1}{S}} \quad (5.19)$$

Onde:

- N: Refere-se ao número de dados;
- F_i : Refere-se ao estimador de frequência seguindo a abordagem de Hosking (1990);
- D_i : Refere-se a diferença entre a Precipitação e PET para o mês i.

A função de distribuição de probabilidade de D de acordo com a distribuição Log-logística ([VICENTE-SERRANO et al., 2010](#)) é então apresentada pela equação 20:

$$F(X) = [1 + (\frac{\alpha}{x - \gamma})^\beta]^{-1} \quad (5.20)$$

Com F(x) o SPEI pode ser facilmente obtido como os valores padronizados de F(x). Por exemplo, seguindo a aproximação clássica de [Abramowitz e Stegun \(1965 apud VICENTE-SERRANO et al., 2010\)](#), como demonstra a equação 21:

$$SPEI = W - \frac{C_0 + C_1W + C_2W^2}{1 + d_1W + d_2w^2 + d_3W^3} \quad (5.21)$$

Onde:

- $W = \sqrt{-2\ln(P)}$ para $P \leq 0.5$;
- C_0 : 2.515517;
- C_1 : 0.802853;
- C_2 : 0.010328;
- d_1 : 1.432788;
- d_2 : 0.189269;
- d_3 : 0.001308.

O SPEI é uma variável padronizada e, portanto, pode ser comparado com outros valores do SPEI no tempo e no espaço ([VICENTE-SERRANO et al., 2010](#)). Por fim, além do SPI, esse índice será parte importante para a criação do texto a fim de verificar a severidade das secas nos últimos 30 anos referente ao território de Brasília-DF. Vale ressaltar que a classificação dos intervalos é feita da mesma maneira como demonstrado na tabela 2.

5.2 Modelagem e inferência

A etapa de modelagem dos dados tem como entregue a escolha e a construção de um modelo preditivo, do qual é avaliado através de testes de hipótese e conceitos estatísticos. Dessa forma, por inferência, o modelo visa responder à pergunta norteadora discutida anteriormente, isto é, a previsão da frequência de incêndios dado as variáveis climáticas.

5.2.1 Regressão linear

Como citado anteriormente, o conceito de regressão linear torna-se necessário para o entendimento do texto, do qual abordará essa técnica matemática visando a previsão de índices de secas dado a variável explicativa precipitação ou evapotranspiração. Portanto, o conceito será abordado de maneira simples para enriquecer o texto posterior.

Os modelos de regressão linear fazem parte de um conjunto de ferramentas comuns entre economistas e estatísticos cujo foco é a realização de inferências, na maior parte das vezes, causais. A inferência consiste em, a partir de evidências encontradas para uma amostra, realizar generalizações de resultados para a população. Ou, de modo mais simples, há um interesse em verificar a correlação entre duas ou mais variáveis e testar o quanto se pode confiar nas estimativas encontradas (CHEIN, 2019).

Portanto, este relacionamento é representado por um modelo matemático, isto é, por uma equação que associa a variável dependente com as variáveis independentes. Este modelo denomina-se como um modelo de regressão linear simples se apenas houver uma relação linear entre a variável dependente e uma variável independente. Se, caso haja mais de uma variável explicativa, o modelo denomina-se como modelo de regressão linear de caráter múltiplo. Entretanto, o texto utilizará apenas a abordagem simples. Por fim, a equação 22 apresenta a fórmula da regressão linear simples.

$$Y = \beta_0 + \beta_1 x + \epsilon \quad (5.22)$$

Onde:

- X: representa a variável explicativa, ou independente, sem erro (não aleatória);
- β_0, β_1 : representam os parâmetros ou coeficientes da regressão a estimar;
- ϵ : representa a variável aleatória residual na qual se visa incluir todas as influências no comportamento da variável de Y que não podem ser explicadas linearmente pelo comportamento da variável X;
- Y: representa a variável resposta, ou dependente (aleatória).

Além disso, existem vários métodos e pressupostos utilizados para validar e verificar a explicabilidade dos parâmetros estimados, dos quais dão origem à reta de regressão, quando considerado tais variáveis resposta e explicativa. Então, o texto utilizará o método dos mínimos quadrados para a estimativa dos parâmetros e as métricas R^2 e Raiz quadrada do erro-médio (RMSE) para a avaliação e validação, além dos cinco pressupostos: linearidade entre as variáveis, os erros provêm de uma distribuição normal, os erros possuem média 0, homoscedasticidade e dos erros não são correlacionados.

5.2.2 Random Forest

O Random Forest Regressor é um algoritmo de aprendizado de máquina que se baseia em um conjunto de árvores de decisão para realizar tarefas de regressão, considerado um modelo ensemble, uma vez que é composto por múltiplas árvores de decisão independentes que trabalham em conjunto para gerar uma previsão final mais precisa e robusta. Essas árvores de decisão são construídas a partir de uma amostra aleatória dos dados de treinamento, e cada árvore é treinada de forma independente das outras. No final, a previsão final é obtida a partir da média das previsões de todas as árvores no conjunto.

Além disso, O Random Forest é um algoritmo de aprendizado de máquina utilizado em tarefas de classificação e regressão. Como citado anteriormente, o algoritmo funciona construindo várias árvores de decisão com diferentes amostras aleatórias dos dados de treinamento e selecionando aleatoriamente os recursos em cada nó da árvore. Para a tarefa de regressão, cada árvore emite uma previsão, e a previsão final é determinada pela média das previsões de todas as árvores no conjunto. Embora a técnica de Random Forest seja mais conhecida por sua aplicação em problemas de classificação, ela também é eficaz em problemas de regressão.

Portanto, a floresta aleatória é um algoritmo de aprendizado supervisionado que usa o método de aprendizado em conjunto para regressão. O método de aprendizado conjunto é uma técnica que combina previsões de vários algoritmos de aprendizado de máquina (combinação de diversas árvores de decisão) para realizar uma previsão mais precisa do que um único modelo ([BREIMAN, 2001](#)).

6 Desenvolvimento do Projeto

6.1 Fluxo de desenvolvimento

Essa seção visa demonstrar o fluxo utilizado para a realização do texto de maneira diagramada, a fim de tornar a percepção dos algoritmos e técnicas estatísticas mais simples sobre o estudo em sua totalidade. Portanto, a seguir é apresentado um fluxograma da qual representa a idealização do texto e suas características.

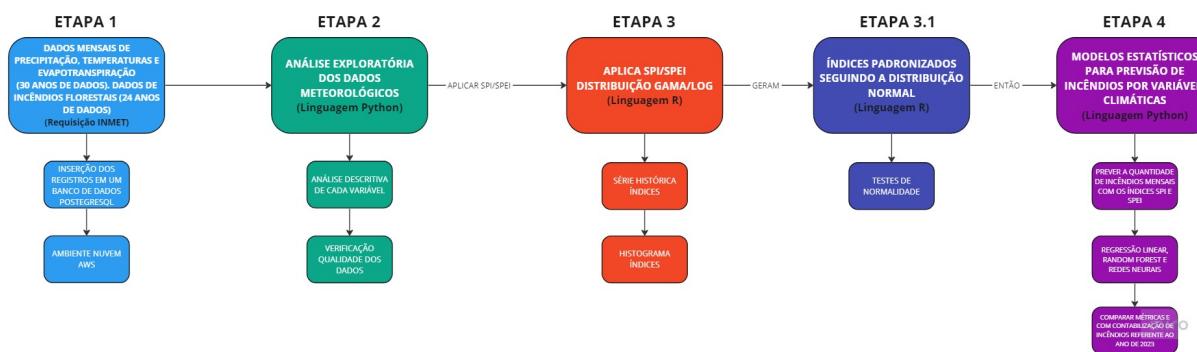


Figura 5 – Fluxo de desenvolvimento

Como é possível verificar na figura acima, o estudo inicia-se a partir da coleta dos dados menais referente ao período de 30 anos para as respectivas variáveis, do qual foi realizada via requisição no site do Instituto Nacional de Meteorologia (INMET), como citado em capítulos anteriores. Também, fez-se a extração de dados referentes a quantidades de incêndios mensais no site do Instituto Nacional de Pesquisas Espaciais (INPE) no período de 24 anos. Além disso, tais dados são inseridos em um banco de dados relacional PostgreSQL no ambiente da nuvem AWS. Vale ressaltar que a inserção de tais dados no ambiente citado foi realizado através da linguagem Python.

A seguir, a análise exploratória dos dados, referente a etapa 2, é estritamente necessária na análise da qualidade dos dados e avaliação estatística de cada uma das variáveis, que serão utilizadas para a composição do texto. Ademais, nessa etapa serão evidenciados valores estatísticos como medidas de tendência central, dispersão, quantidade de dados nulos e dados discrepantes, dos quais enriquecerão as etapas de modelagem posteriores. Por fim, essa será realizada com a utilização da linguagem Python e suas respectivas bibliotecas voltadas para a manipulação e visualização dos dados.

Em relação à etapa 3, decidiu-se pela divisão em duas, onde a primeira é responsável pela aplicação dos índices SPI e SPEI por meio da linguagem R, da qual conta com uma biblioteca referência para o cálculo de tais índices. Além disso, serão avaliadas as séries históricas e histogramas oriundas dos índices gerados com os dados meteorológicos.

Já a etapa 3.1 contará com os valores gerados após a aplicação dos índices SPI e SPEI. Então, essa fase fará a validação dos índices, dos quais devem seguir uma distribuição normal, como citado na etapa de referencial teórico. Portanto, por meio da linguagem R será validado através de testes estatísticos voltados à normalidade dos dados oriundos dos índices gerados.

Por fim, a etapa 4 contará com a criação de modelos estatísticos dos quais têm como objetivo relacionar a severidade dos índices de secas com a contabilização dos focos de incêndios. Para isso, decidiu-se pela aplicação de três modelos: regressão linear, random forest e redes neurais. Portanto, essa etapa tentará prever a quantidade de incêndios da respectiva localidade ao considerar os índices de estiagem previamente calculados na seção 3.1 e comparar com dados atuais do ano de 2023. Vale ressaltar que essa etapa será realizada por meio de bibliotecas de manipulação e modelagem de dados presentes na linguagem Python. Além disso, será implementada uma aplicação web para tornar os resultados interativos com os possíveis usuários que se interessarem sobre.

6.2 Análise Exploratória dos Dados

A análise exploratória visa entender e investigar os dados, principalmente numérica e estatisticamente, buscando entender como estão associados as possíveis distribuições para assim extrair informações úteis das próximas etapas do texto. Para a análise a seguir utilizaram-se 372 registros mensais referentes a dados meteorológicos em um intervalo de 30 anos, isto é, nos anos de 1992 a 2022. Além disso, é válido ressaltar que a análise exploratória foi efetuada utilizando a linguagem Python, como exposto no fluxograma de desenvolvimento.

6.2.1 Precipitação mensal

Como informado anteriormente, a precipitação refere-se a um fenômeno natural que ocorre quando a água cai da atmosfera e atinge a superfície da Terra na forma líquida ou sólida. A quantidade de precipitação que ocorre em uma área pode ser medida em milímetros ou polegadas e influenciada por fatores como a temperatura, umidade, altitude, topografia e a presença de massas de ar quente ou frio. Além disso, a precipitação é importante para a agricultura, hidrologia, meteorologia e outras áreas da ciência tornando-se vital para o funcionamento dos ecossistemas terrestres. No entanto, esse fenômeno quando excessivo ou insuficiente pode ter impactos negativos na sociedade e no meio ambiente.

A fim de verificar como os dados referentes a precipitação se distribuí, fez-se necessário a aplicação de técnicas estatísticas para a avaliação, principalmente por via de gráficos com as visualizações de tais registros. Portanto, de maneira inicial, decidiu-se

verificar a série histórica referente a frequência mensal acumulada de precipitação nos últimos 30 anos referente ao território de Brasília-DF, do qual possui 372 registros e verifica-se abaixo.

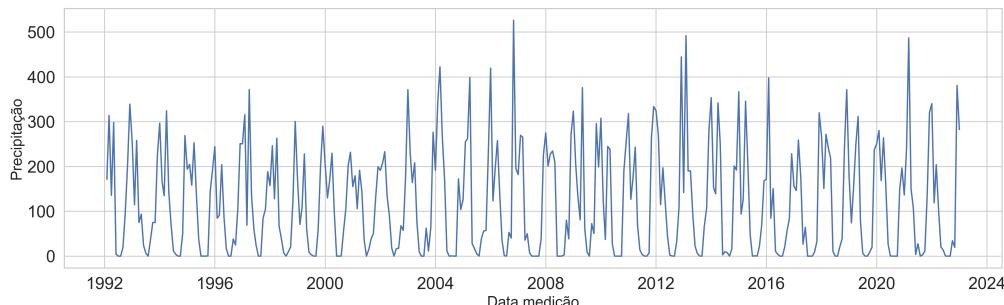


Figura 6 – Série histórica - Precipitação

Conforme a figura 6, é possível comprovar a série temporal referente aos dados de precipitação no período citado. Dessa forma, verifica-se ciclos e sazonalidade na série, com picos e vales, uma vez que se trata de dados meteorológicos, dos quais são afetados por estações climáticas. Ademais, não é possível identificar uma tendência da série, característica oriunda de dados climáticos. Por fim, verifica-se que os três maiores picos de precipitação ocorreram após a década de 1990, principalmente a partir do ano de 2005.

Logo, ao verificar as estatísticas básicas, referentes às medidas de tendência central e dispersão, identificou-se que a média mensal de precipitação nos 30 anos de coleta dos dados foi igual a 122 milímetros. Da mesma maneira, a precipitação mensal máxima foi igual a 526.4 milímetros em outubro de 2004. Entretanto, no período citado, houveram 57 meses dos quais a precipitação foi de 0 milímetros. Ademais, confirma-se que pelo menos metade dos meses possuíram precipitação mensal igual ou inferior a 92.4 milímetros. Por fim, abaixo é apresentado a distribuição da frequência de precipitação mensal por meio de um gráfico de histograma.

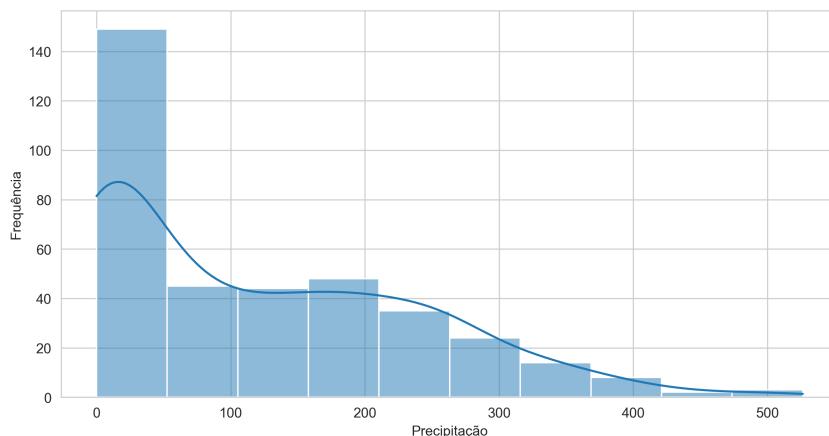


Figura 7 – Histograma - Precipitação

O principal propósito do histograma é verificar a distribuição das frequências de dados contínuos considerando as classes. Dessa forma, por meio da alta frequência da primeira classe, e da característica assimétrica à direita, na figura 7, torna-se evidente o clima semi-árido presente na maior parte do ano no território de Brasília-DF, do qual possui o cerrado como bioma.

Por fim, dado as características encontradas no histograma: assimetria à direita e cauda longa, a distribuição estatística mais comumente utilizada para modelar dados de precipitação é a distribuição gama. Então, essa distribuição é adequada para modelar dados de precipitação devido às suas propriedades, como assimetria positiva (ou à direita) e cauda longa, que refletem a distribuição real da precipitação. Dessa forma, aplicou-se o teste de Kolmogorov-Smirnov a fim de verificar se tais dados seguem a respectiva distribuição, resultando em um p-valor de 0.2, do qual é acima do nível de significância de 0.05, comprovando que os registros de precipitação seguem uma distribuição gama.

6.2.2 Temperatura máxima média mensal

Os dados referentes a temperatura máxima média coletada mensalmente no período de 30 anos citado, ajudará para a criação da temperatura média mensal, da qual será analisada posteriormente. Portanto, de maneira inicial, fez-se a análise de séries temporais referente a tal variável, como é apresentado abaixo.

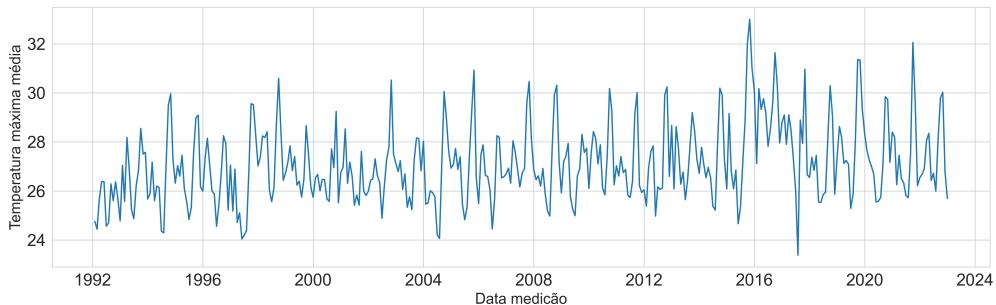


Figura 8 – Série histórica - Temperatura máxima média

Conforme a figura 8, é possível comprovar a série temporal referente aos dados de temperatura máxima média citado. Dessa forma, não é possível verificar-se de maneira tão clara os ciclos e sazonalidade na série. Ademais, é possível identificar uma leve tendência positiva da série, demonstrando que, com o passar dos anos, a temperatura máxima média mensal está se tornando ligeiramente maior. Consequentemente, verificam-se picos, principalmente a partir do ano de 2015, dos quais, permite afirmar uma grande amplitude climática durante os anos.

Logo, ao verificar as estatísticas básicas, referentes às medidas de tendência central e dispersão, identificou-se que a temperatura máxima mensal foi igual a 33 graus Celsius

em outubro de 2015, número do qual é considerado extremamente alto em tal cenário. Já a temperatura máxima mais baixa registrada foi de 23.3 graus Celsius em julho de 2017. Ademais, confirma-se que pelo menos metade dos meses possuíram temperatura máxima média mensal igual ou inferior a 26.8 graus Celsius. Por fim, abaixo é apresentado a distribuição da frequência de tal variável por meio de um gráfico de histograma.

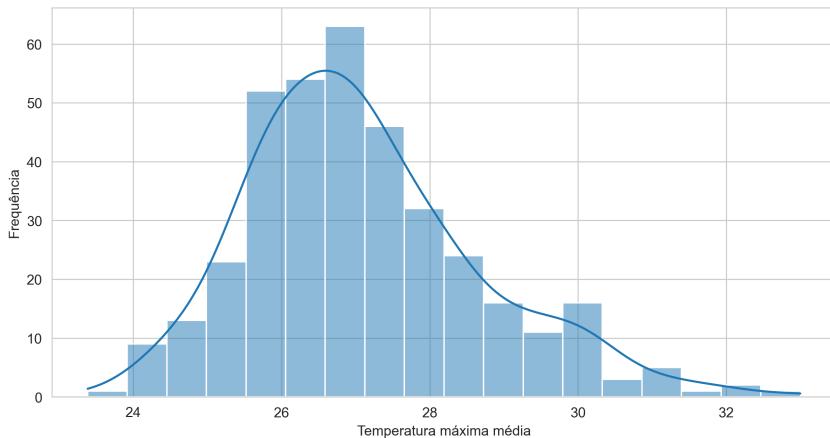


Figura 9 – Histograma - Temperatura máxima média

Dessa forma, a figura 9 apresenta a distribuição da frequência referente aos dados de temperatura máxima mensais aparentemente em um formato encontrado na distribuição normal. Dessa forma, aplicou-se o teste Shapiro-Wilk para verificar a normalidade de tais dados, resultando em um p-valor igual a 0.00000016, do qual é abaixo do nível de significância adotado de 0.05, comprovando que tais dados não seguem uma distribuição normal.

6.2.3 Temperatura mínima média mensal

Os dados referentes a temperatura mínima média coletada mensalmente no período de 30 anos citado, ajudará para a criação da temperatura média mensal, da qual, como explicado no tópico anterior, será analisada posteriormente. Portanto, de maneira inicial, fez-se a análise de séries temporais referente a tal variável, conforme apresentado abaixo.

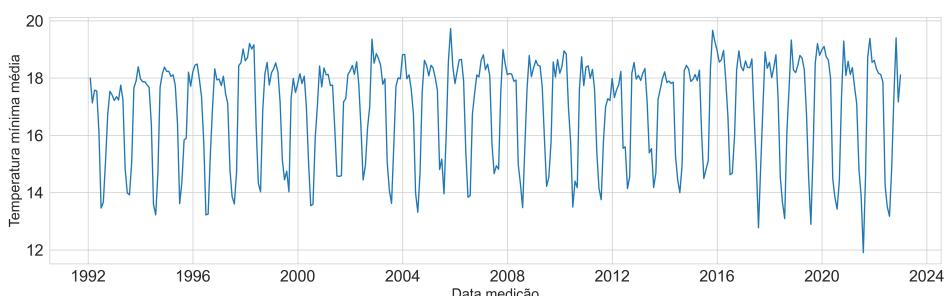


Figura 10 – Série histórica - Temperatura mínima média

De maneira inicial, diferente do que foi apresentada na série temporal referente às temperaturas máximas, para a variável agora de análise verifica-se uma amplitude térmica mais controlada. Da mesma forma, a série histórica referente a temperatura mínima média mensal apresenta claros ciclos e sazonalidade, mas não sendo possível determinar uma tendência.

Logo, ao verificar as estatísticas básicas, referentes às medidas de tendência central e dispersão, identificou-se que a temperatura máxima mensal foi igual a 19.7 graus Celsius em outubro de 2021. Já a temperatura máxima mais baixa registrada foi de 11.9 graus Celsius em julho de 2017. Ademais, confirma-se que pelo menos metade dos meses possuíram temperatura mínima média mensal igual ou inferior a 17.6 graus Celsius no período entre 1992 a 2022. Por fim, abaixo é apresentado a distribuição da frequência de tal variável por meio de um gráfico de histograma.

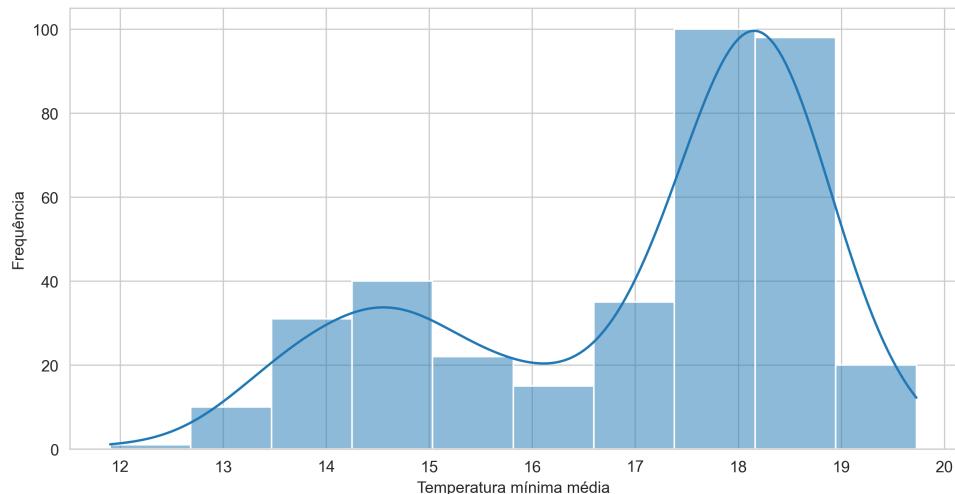


Figura 11 – Histograma - Temperatura mínima média

Dessa forma, a figura 11 apresenta a distribuição da frequência referente aos dados de temperatura máxima mensais. Entretanto, decidiu-se testar a normalidade dos dados, uma vez que a variável referente a temperatura máxima possuiu tal comportamento. Dessa forma, aplicou-se o teste Shapiro-Wilk para verificar a normalidade de tais dados, resultando em um p-valor igual a 0.00000000000000104, do qual é abaixo do nível de significância adotado de 0.05. Além disso, definiu-se a utilização de outro teste de normalidade a fim de fortalecer o resultado anterior. Então, aplicou-se o teste D'Agostino-Pearson, resultando em um p-valor igual a 0.0000000000019, do qual é abaixo do nível de significância adotado de 0.05 comprovando que tais dados não seguem uma distribuição normal.

6.2.4 Temperatura média mensal

Os dados a respeito da temperatura média mensal são essenciais visando o texto a seguir, uma vez dos quais serão utilizados os índices climáticos propostos inicialmente:

SPI e SPEI. Da mesma forma, para o cálculo, utilizou-se a média aritmética comum, isto é, a partir das respectivas temperaturas máxima e mínimas mensais. Portanto, abaixo é apresentado a série histórica referente aos dados de temperatura média mensal no período citado.

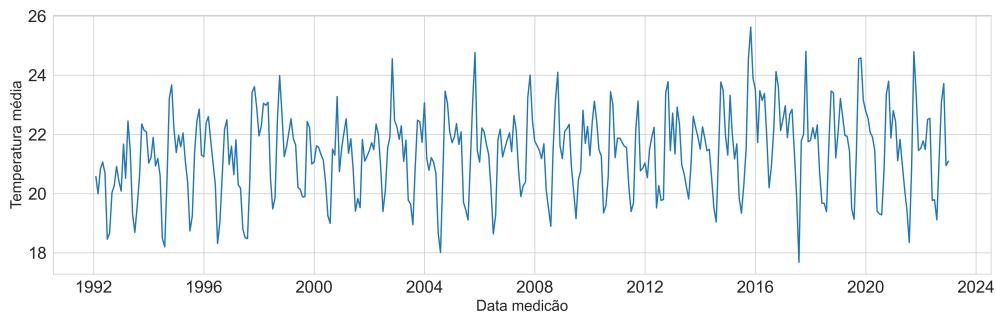


Figura 12 – Série histórica - Temperatura média mensal

Conforme a figura 12, é possível identificar uma leve tendência positiva referente a ascensão da temperatura média com o passar dos anos, fenômeno do qual foi identificado também na figura 8. Além disso, verifica-se uma grande amplitude térmica, uma vez que a menor temperatura média mensal registrada foi de 17.6 graus Celsius em julho de 2017, já a maior temperatura média mensal registrada foi de 25.6 graus Celsius em outubro de 2015. Ademais, confirma-se que pelo menos metade dos meses possuíram temperatura mínima média mensal igual ou inferior a 21.4 graus Celsius no período entre 1992 a 2022. Por fim, abaixo é apresentado o histograma referente a variável de interesse.

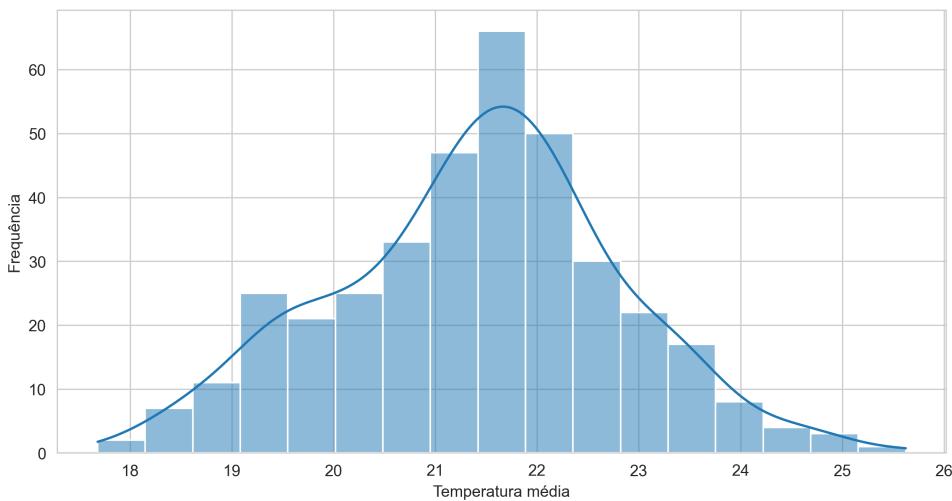


Figura 13 – Histograma - Temperatura média

Dessa forma, a figura 13 apresenta a distribuição da frequência referente aos dados de temperatura média mensais supostamente em um formato encontrado na distribuição normal. Dessa forma, aplicou-se o teste Shapiro-Wilk para a verificação da normalidade

referente a tais dados, resultando em um p-valor igual a 0.099, do qual é acima do nível de significância adotado de 0.05, comprovando que tais dados seguem uma distribuição normal.

6.2.5 Incêndios florestais

A exploração dos dados referentes à série histórica sobre incêndios florestais torna-se necessária, uma vez que esse estudo tem o intuito de relacionar os índices de seca posteriormente calculados com o histórico aqui explorado. Dessa forma, abaixo é possível identificar a série histórica com as respectivas frequências de incêndio por ano.

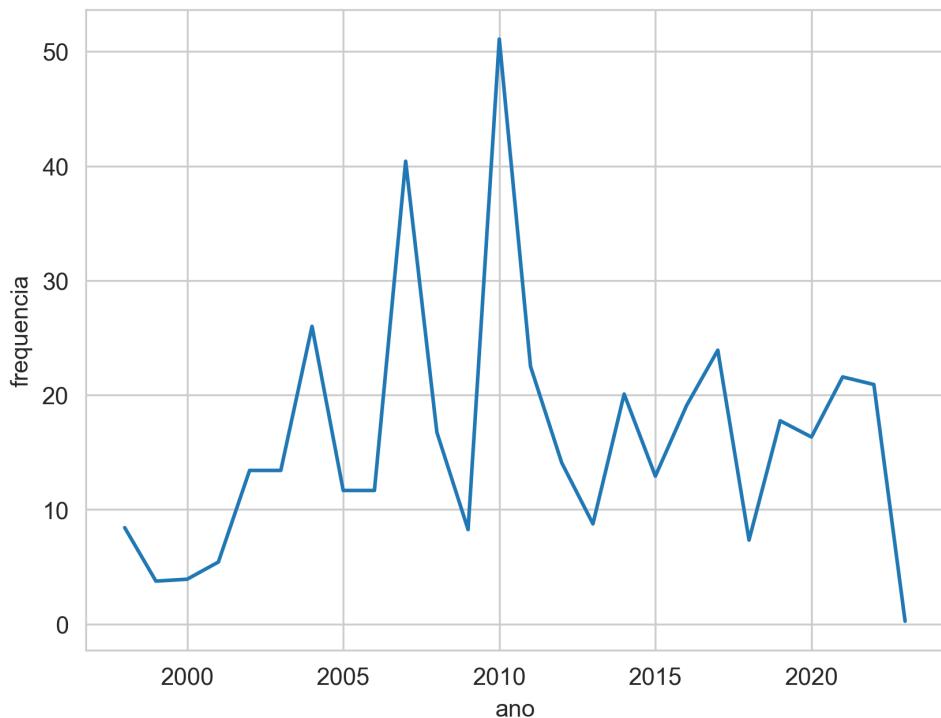


Figura 14 – Série histórica - Incêndios florestais

A figura 14 demonstra o quantitativo de incêndios ocorridos na localidade de Brasília-DF referente ao período de 1998 a 2022. Portanto, é possível identificar que os anos de 2007 e 2010 foram os representantes dos picos evidenciados, com 485 e 613 focos de incêndios, respectivamente. Da mesma maneira, entende-se que os demais anos seguiram variações não discrepantes se comparados entre si. Por fim, abaixo é possível perceber a ocorrência de incêndios nos meses que a seca alcança uma intensidade maior em Brasília-DF, além da respectiva sazonalidade, uma vez que esse comportamento está relacionado a eventos que sempre acontecem em uma determinada época.

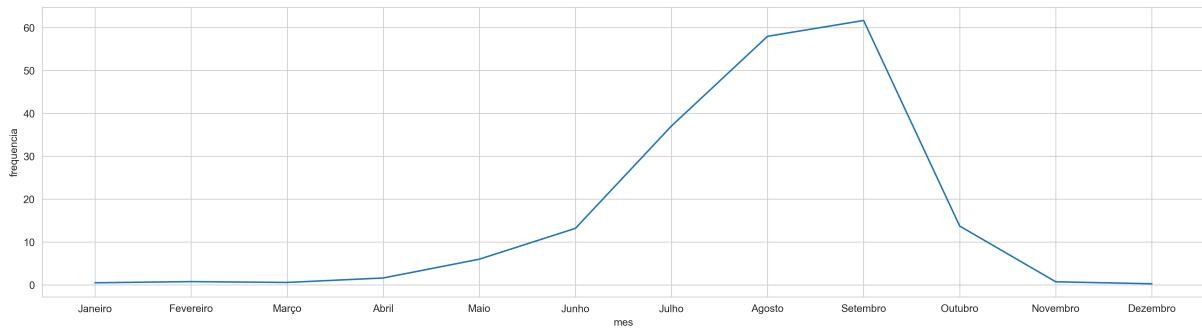


Figura 15 – Série histórica mensal - Incêndios florestais

6.2.6 Relação precipitação e frequência de incêndios

Como demonstrado anteriormente, foi possível entender que os meses com mais frequência de queimadas referem-se aos meses de seca no inverno de Brasília-DF, englobando os meses de maio a outubro. Dessa forma, a precipitação e a quantidade de incêndios podem estar diretamente relacionadas em Brasília-DF, uma vez que, durante essa época, a quantidade de chuvas diminui significativamente, resultando em uma redução na umidade do solo e na vegetação.

À vista disso, a escassez de chuvas torna o ambiente mais propenso a incêndios, uma vez que a vegetação se torna seca e inflamável. Além disso, a falta de umidade dificulta o controle e a extinção de incêndios, pois não há a presença de água suficiente para combater as chamas. Portanto, abaixo é possível evidenciar tal fato por meio do gráfico relacionando a quantidade de incêndios e precipitação.

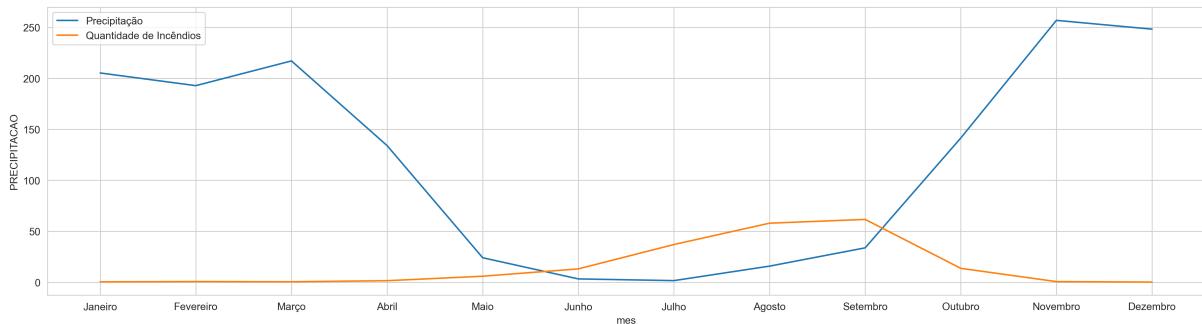


Figura 16 – Relacionamento Precipitação e Queimadas

Então, a figura acima demonstra que, nos meses em que há o crescimento de incêndios florestais, a quantidade de precipitação abaixa significantemente (em certos casos até em 0 mm), fenômeno do qual se inverte após os meses de seca na localidade. Dessa forma, tal gráfico confirma a informação referente à relação entre as duas variáveis.

6.2.7 Correlação

A correlação é uma das principais técnicas estatísticas para a verificação do relacionamento entre pares de variáveis quantitativas. É valido ressaltar que, a correlação utilizada no texto será a de Pearson, do qual, de maneira simples, possui o intervalo de -1 a 1. Dessa forma, quando igual a -1 a correlação é linearmente negativa, quando igual a 1 a correlação é linearmente positiva e se igual a 0, a correlação é nula.

O principal objetivo dessa análise é entender como as variáveis de temperatura podem afetar a precipitação e vice-versa. Portanto, após a computação de tal métrica, gerou-se a matriz de correlação citada anteriormente, da qual pode ser visualizada abaixo.

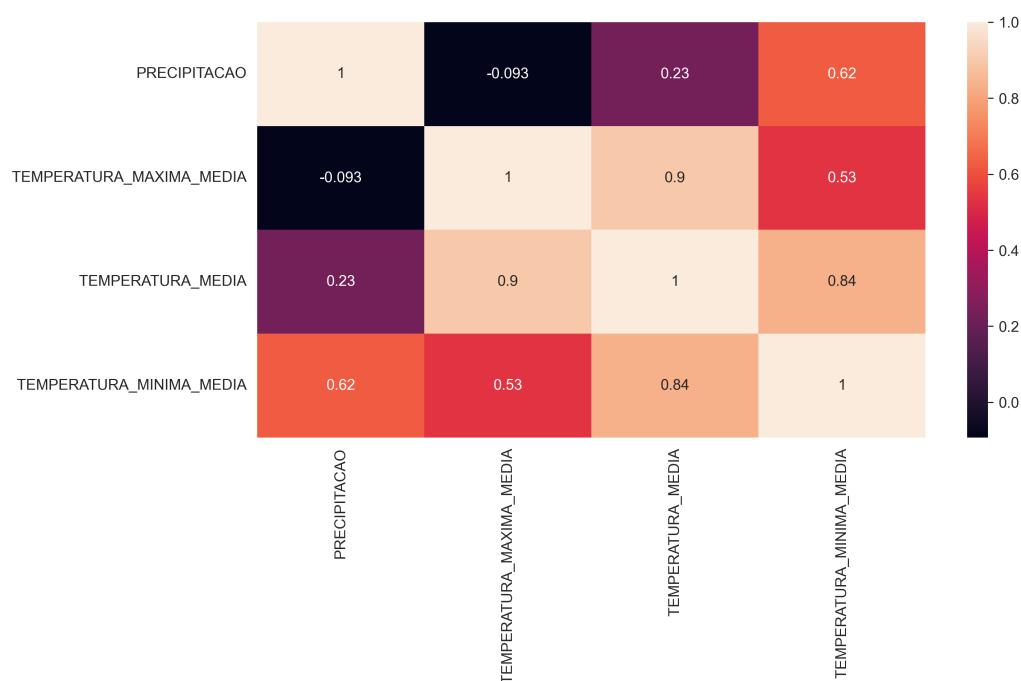


Figura 17 – Matriz de correlação

Conforme a figura 17, da qual apresenta a matriz de correlação de Pearson, é possível identificar que as variáveis climáticas referentes a temperatura afetam a precipitação de maneira linear. Dessa forma, verificou-se que a temperatura mínima média mensal possuiu uma correlação moderadamente positiva em relação à precipitação, ocasionando em uma relação linear entre o par, fato do qual possui coerência, uma vez que é comum temperaturas mais baixas serem em dias chuvosos. Além disso, a temperatura máxima média mensal possuiu uma correlação fortemente negativa se comparado à precipitação, informação do qual possui nexo, uma vez que dias quentes não costumam chover na região do cerrado, do qual é o bioma presente na localidade que tais dados foram coletados.

6.2.8 Qualidade dos dados

A análise sobre a qualidade dos dados é essencial para garantir a confiabilidade e a precisão das informações utilizadas em processos decisórios. Dados de baixa qualidade podem levar a resultados imprecisos, erros e prejuízos financeiros. Além disso, a falta de confiança nos dados pode levar à perda de credibilidade da empresa ou organização. A análise da qualidade dos dados também é importante para identificar problemas, como dados duplicados, ausência de informações ou inconsistências, permitindo que sejam corrigidos antes de prejudicar o processo de tomada de decisão. Portanto, investir na análise da qualidade dos dados é fundamental para garantir a eficácia e a eficiência nos processos de gestão e na obtenção de resultados positivos. Dessa forma, a tabela a seguir demonstra a primeira etapa da análise, quando decide-se avaliar os dados faltantes para cada variável.

Tabela 3 – Quantidade de dados nulos por variável

Variável	Quantidade nulos
Data medição	0
Precipitação	0
Temperatura máxima média	0
Temperatura mínima média	0
Temperatura média	0
Frequência de incêndios	0

Fonte: autor, 2023

O gráfico de boxplot é uma ferramenta gráfica muito útil para representar e visualizar a distribuição de dados, permitindo identificar possíveis pontos discrepantes (outliers) ou valores extremos. Essa técnica utiliza uma caixa retangular para representar a distribuição dos dados, sendo que a mediana é representada por uma linha na caixa e os quartis são representados pelas bordas da caixa. Consequentemente, os pontos discrepantes são representados por pontos fora da caixa, representando registros atípicos em relação à distribuição ordinária.

Logo, ao utilizar o gráfico de boxplot, é possível identificar rapidamente se existem valores extremos que podem afetar a análise dos dados. Isso ocorre porque os pontos discrepantes são representados de forma visualmente distinta dos demais dados, permitindo que sejam facilmente identificados. Consequentemente, identificar os dados discrepantes é importante porque eles podem ser indicativos de erros nos dados, como, por exemplo, dados coletados incorretamente ou problemas na entrada de dados, uma vez que a coleta de tais registros são oriundos de uma estação convencional. Abaixo é possível verificar o gráfico de boxplot para cada uma das variáveis das quais serão utilizadas para a criação desse estudo.

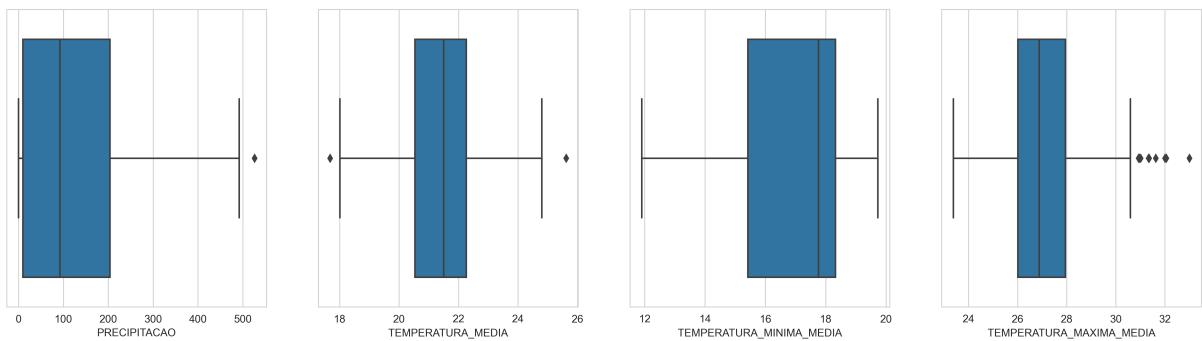


Figura 18 – Boxplot das variáveis de estudo

Conforme a figura 18 apresenta, as variáveis apresentam pelo menos um ponto discrepante, com exceção da temperatura mínima média mensal. Dessa forma, a temperatura máxima média mensal possuiu cinco registros dessa natureza, já a temperatura média mensal apenas dois dados destoantes em relação aos demais, e a precipitação apenas um único registro do qual não segue a distribuição imposta para tal amostra. Portanto, após a análise da qualidade dos dados, verifica-se que os dados utilizados para o texto proposto possuem grande confiabilidade com relação a sua completude, confiança e veracidade.

6.3 SPI

O Índice de Precipitação Padronizada (SPI) é uma medida estatística amplamente utilizada para avaliar a seca em uma região específica. É um indicador que considera a quantidade e a distribuição da precipitação em uma determinada série histórica, e pode ser calculado em diferentes escalas, variando de semanas a décadas. Assim sendo, o SPI é útil para a prevenção e previsão de secas, por permitir monitorar as condições meteorológicas de uma região e avaliar se a precipitação está dentro do esperado para o período em questão. Além disso, o índice SPI também pode ser utilizado para prever secas futuras, com base em análises estatísticas das séries históricas ao considerar dados de precipitação. Portanto, é possível antecipar situações de escassez de água e tomar medidas preventivas para minimizar seus impactos negativos na economia, na agricultura e na população em geral.

Então, a análise a seguir visa demonstrar a implementação, na linguagem de programação R, do cálculo e visualização do índice SPI6, avaliando a seca em escalas de tempo de 6 meses por conta do histórico da localidade escolhida, uma vez que a escala é um dos parâmetros mais importantes na modelagem dos índices de seca SPI e SPEI, podendo ser ajustada conforme a resolução dos dados disponíveis e a variabilidade da seca na região em questão. Portanto, abaixo é apresentado a série histórica do índice, do qual pode ser interpretada com a tabela 2.

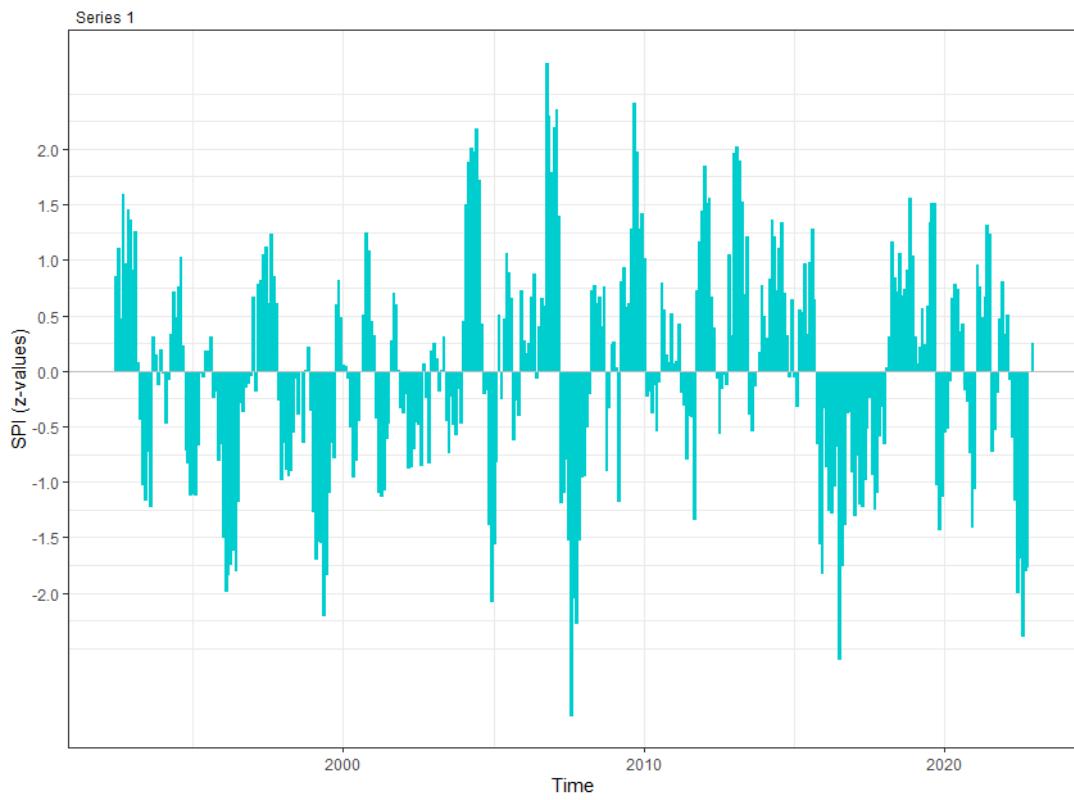


Figura 19 – Série histórica índice SPI6

Na figura acima é apresentada a série temporal com o índice SPI6 referente aos dados de 1992 a 2022 da localidade de Brasília-DF. Consequentemente, ao avaliar a tabela 2, entende-se que todos os pontos dos quais estão abaixo do índice -1 correspondem a um clima muito seco, englobando diversos anos. Além disso, percebem-se certos picos dos quais avançam o valor de -2, considerando tais dados como extremamente secos. Entretanto, no ano de 2017 é possível identificar um período extenso contendo apenas índices negativos, do qual corresponde à realidade de Brasília-DF em tal ano, uma vez que se foi necessário políticas de racionamento de água, dada a estiagem que afetava a cidade. Por fim, é possível entender uma despadronização quanto a variabilidade do índice durante o período estudado.

Como explicado anteriormente, o índice SPI segue uma distribuição normal. Entretanto, a fim de verificar tal fenômeno, decidiu-se pela realização de um teste de normalidade sobre os dados referente aos índices gerados. Dessa forma, com um nível de significância (α) a 5%, formularam-se as seguintes hipóteses:

H_0 : Os índices seguem uma distribuição normal

H_1 : Os índices não seguem uma distribuição normal

Após a aplicação do teste de Shapiro-Wilk por meio da linguagem de programação R, gerou-se o seguinte histograma do qual pode-se concluir que os índices seguem uma

distribuição normal, aceitando a hipótese nula (H_0), uma vez que o p-valor encontrado, de 0.991, encontra-se acima do nível de significância adotado, de 0.05.

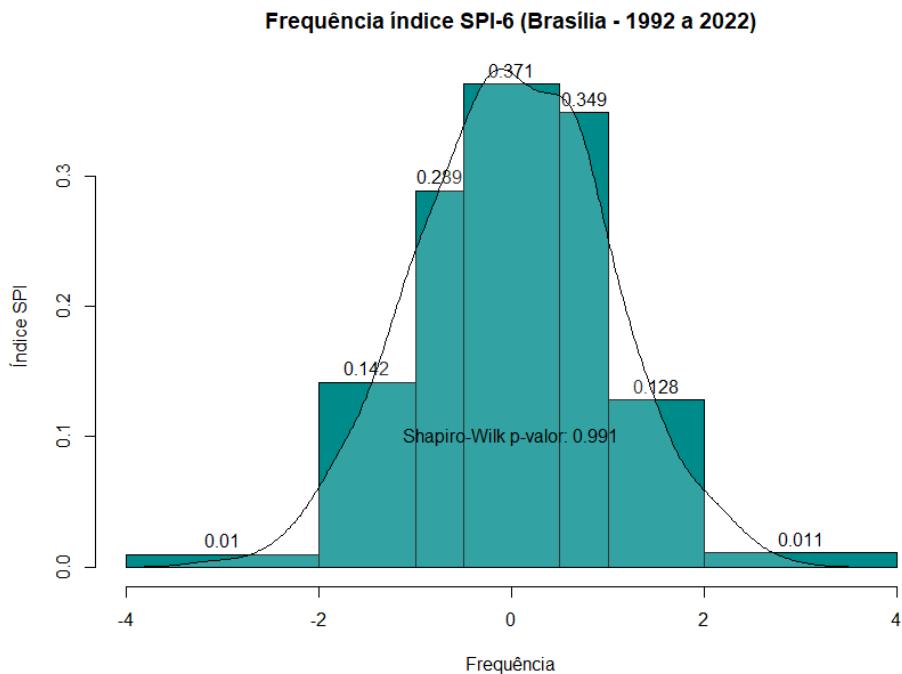


Figura 20 – Distribuição índices SPI6

Portanto, é possível identificar no histograma da figura 20 a frequência de cada classe, da qual corresponde ao respectivo grau de severidade de seca. A classificação normal (0.49 a -0.49) foi a mais frequente nos dados analisados e assim por diante, uma vez que houve a comprovação que tais dados seguem uma distribuição normal, em formato de sino.

6.4 SPEI

O Índice Padronizado de Evapotranspiração (SPEI) é uma medida que, assim como o SPI, também é utilizada para avaliar a seca em uma determinada região. A principal diferença entre os dois índices é de que o SPEI considera tanto a precipitação quanto a evapotranspiração, ou seja, a quantidade de água que evapora da superfície do solo e das plantas, conhecido como balanço hídrico. Por isso, a literatura considera o SPEI mais eficiente que o SPI para descrever eventos secos e úmidos em períodos longos, já que a evapotranspiração é um processo que afeta diretamente a disponibilidade de água no solo e nas plantas. Dessa forma, ao considerar a evapotranspiração, o SPEI pode fornecer uma melhor estimativa das condições climáticas futuras e, consequentemente, das condições de seca.

É válido relembrar os pontos discrepantes encontrados na etapa de qualidade dos dados (seção 6.2.8) referentes a variável de temperatura máxima média mensal.

Consequentemente, como citado anteriormente, o índice SPEI é influenciado por diversos fatores climáticos, incluindo temperatura e precipitação. Portanto, temperaturas altas podem aumentar a evapotranspiração. Dessa forma, o SPEI, por sua vez, tenta medir a umidade do solo e a disponibilidade de água, considerando os efeitos de longo prazo da evapotranspiração e da precipitação. Por fim, decidiu-se filtrar tais pontos discrepantes, uma vez que filtrar temperaturas acima de um determinado valor podem melhorar o modelo SPEI. Além disso, como citado no referencial teórico, o cálculo do índice SPEI considera a Balanço Hídrico Climático (BAL), do qual é obtida através da diferença entre os registros de Precipitação Potencial de Evapotranspiração (PET).

Então, da mesma forma como abordado na análise do índice SPI6, o texto visa demonstrar a implementação, em linguagem R, do cálculo e visualização referente ao índice SPEI3, do qual se decidiu pela escala menor ao considerar o histórico de evapotranspiração da localidade. À vista disto, cada índice (SPI e SPEI) pode requerer uma escala diferente, dependendo das características do clima e da região em que se está avaliando. No entanto, é importante notar que o SPEI é influenciado pela umidade, e, portanto, ser mais sensível a mudanças na escala. Por exemplo, uma escala menor pode levar a uma detecção mais precisa de eventos de seca, porém resultar em maior variabilidade espacial nos valores do índice. Portanto, abaixo é apresentado a série histórica do índice considerando a agregação a cada 3 meses, do qual pode ser interpretada com a tabela 2.

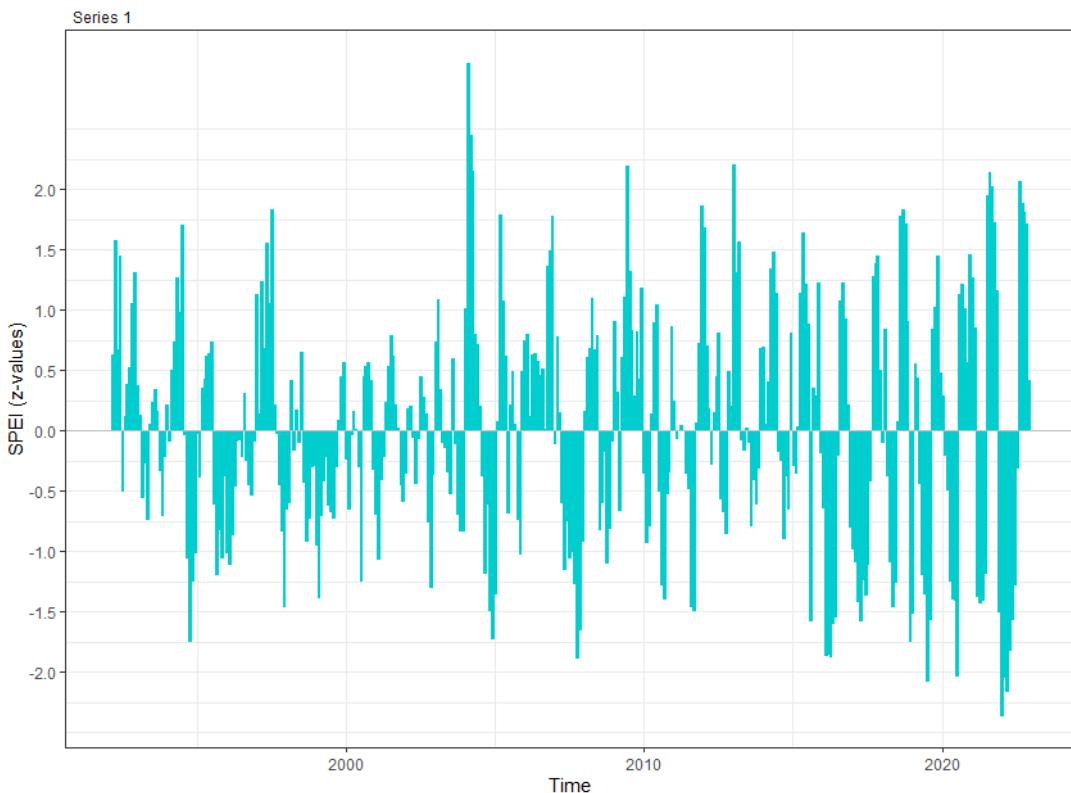


Figura 21 – Série histórica índice SPEI3

Como citado, o índice SPEI aparenta ser mais variado ao considerar a localidade da qual os dados foram extraídos e a escala utilizada, uma vez que Brasília-DF possui um período de seca, de maio a setembro, mas não necessariamente longos períodos com umidade baixa, motivo do qual se escolheu a escala 3 para o índice. Dessa forma, ao avaliar a figura 21, é possível identificar diversos períodos de seca, dos quais são representados abaixo do índice -1 no eixo Y do gráfico, conforme a tabela 2. Além disso, é perceptível períodos com variedades extremas após a metade da década de 2010, o qual pode ser um efeito da escala escolhida, uma vez que escalas menores podem ser mais precisas, porém também acabar resultando em maior variabilidade espacial nos valores do índice.

Da mesma maneira, como aplicado nos índices SPI, essa seção também fará a mesma análise de normalidade dos índices, dessa vez para os índices SPEI. Então, considerando um nível de significância (α) a 5%, formularam-se as seguintes hipóteses:

H_0 : Os índices seguem uma distribuição normal

H_1 : Os índices não seguem uma distribuição normal

Portanto, após a aplicação do teste de Shapiro-Wilk por meio da linguagem de programação R, gerou-se o seguinte histograma do qual pode-se concluir que os índices seguem uma distribuição normal, aceitando a hipótese nula (H_0), uma vez que o p-valor encontrado, de 0.2545, encontra-se acima do nível de significância adotado, de 0.05.

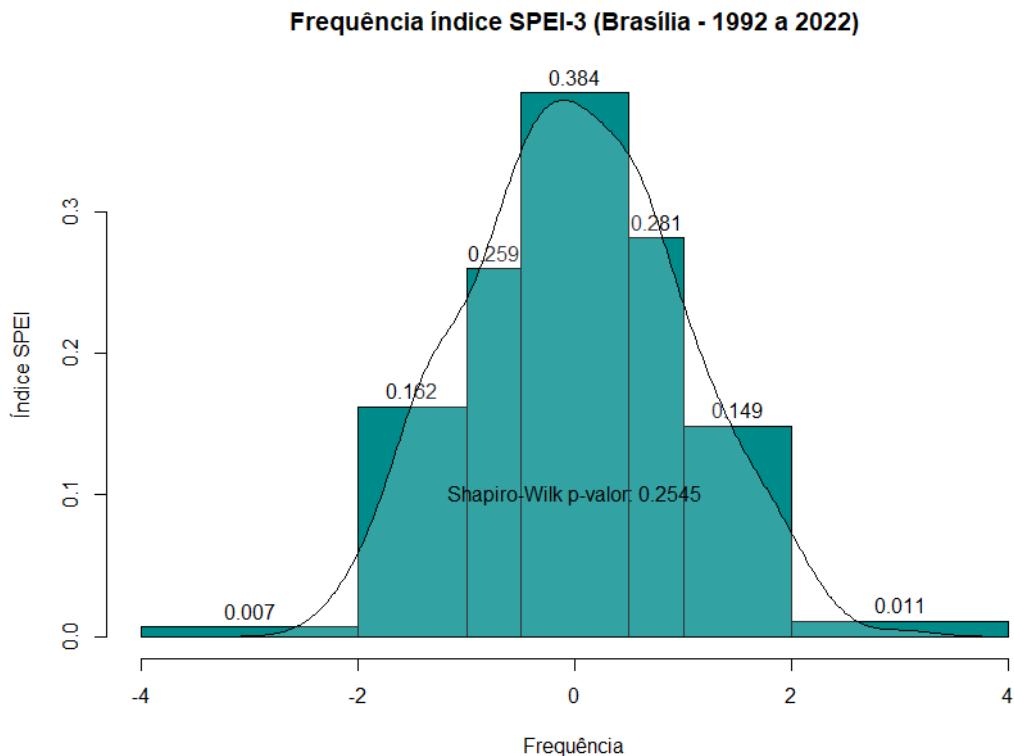


Figura 22 – Distribuição índices SPEI3

Assim, conforme a figura 22 é possível identificar que o histograma gerado possui um comportamento identificado por uma distribuição normal, do qual as respectivas frequências das classes são referentes a classificação disponível na tabela 2. Portanto, como citado no início da sessão, o índice SPEI considera tanto a precipitação quanto a evapotranspiração, ajudando em análises mais assertivas para climas predominantemente secos e com baixa umidade.

6.5 Modelagem dos dados

Determinados autores demonstram uma avaliação similar à proposta a seguir no texto, do qual tenta relacionar linearmente o índice SPI aos incêndios florestais (OLIVEIRA et al., 2017). Da mesma forma, conforme a sessão de planejamento da análise, decidiu-se pela implementação de três modelos estatísticos a fim de prever a frequência de incêndios por meio das variáveis das quais serão selecionadas nessa seção. A seguir será apresentada a modelagem de cada um dos modelos: regressão linear, random forest e redes neurais, em seguida das respectivas análises de métrica, acurácia e pressupostos a fim de validá-los e verificar qual modelagem melhor explica a predição desejada. Vale ressaltar que, como citado no fluxo de desenvolvimento, essa etapa fez-se o uso da linguagem Python e suas respectivas bibliotecas para a implementação da modelagem e as respectivas manipulações de tabelas das quais foram necessárias.

6.5.1 Preparação dos dados

A preparação dos dados é um passo fundamental na implementação de modelos preditivos, pois os dados brutos geralmente são inconsistentes, incompletos ou irrelevantes. Também, essa etapa inclui a integração dos dados para garantir que eles estejam prontos para a análise. O negligenciamento dessa etapa pode levar a modelos imprecisos ou enviesados, prejudicando a tomada de decisões e impactar negativamente os negócios. Portanto, a preparação dos dados é uma etapa crítica que deve ser cuidadosamente realizada antes da implementação dos modelos preditivos.

Então, como dito anteriormente, fez-se necessário unir as duas bases de dados utilizadas nesse estudo até então: a base de dados referente aos registros de seca, já contendo os índices SPI6 e SPEI3 calculados, e o conjunto dos dados referentes a incêndios florestais. Entretanto, precisou-se utilizar uma janela de intersecção para a junção das duas base de dados, uma vez que a primeira referia-se ao período de 1992 a 2022 e a segunda ao período de 1998 a 2022. Portanto, os dados dos quais serão utilizados para a criação e treinamento do modelo faz uso dos registros relacionados mensalmente referente as duas bases de dados com o período de 1998 a 2022.

6.5.2 Seleção das variáveis

A seleção de variáveis é uma etapa fundamental na implementação de modelos preditivos, uma vez que permite identificar quais variáveis são mais relevantes para o modelo, e então descartar aquelas que têm pouco ou nenhum impacto. Da mesma forma, essa etapa ajuda a melhorar a precisão dos modelos e reduzir a complexidade, o que pode levar a modelos mais rápidos e mais simples de interpretar. Além disso, a seleção de variáveis ajuda a reduzir o risco de overfitting, que ocorre quando o modelo é muito complexo e se ajusta demais aos dados de treinamento, prejudicando sua capacidade de generalização. A seleção de variáveis também ajuda a identificar quais variáveis estão correlacionadas, o que pode levar a um modelo mais robusto. Portanto, a seleção de variáveis é uma etapa crítica que deve ser cuidadosamente realizada antes da implementação dos modelos preditivos.

No contexto da seleção de variáveis, a correlação pode ser usada para identificar quais variáveis têm uma forte relação com a variável de saída, isto é, aquela que queremos prever. Isso ajuda a reduzir o número de variáveis a serem consideradas no modelo, o que consequentemente leva a modelos mais rápidos e mais precisos. Além disso, a correlação contribui para identificar potenciais problemas de multicolinearidade, que ocorrem quando duas ou mais variáveis estão altamente correlacionadas entre si, resultando em uma redução na precisão do modelo. Portanto, a correlação pode ser uma ferramenta útil para a seleção de variáveis em modelos preditivos nesse contexto. A seguir é apresentada a matriz e seus respectivos valores para cada par de variáveis referente à base de dados preparada anteriormente.

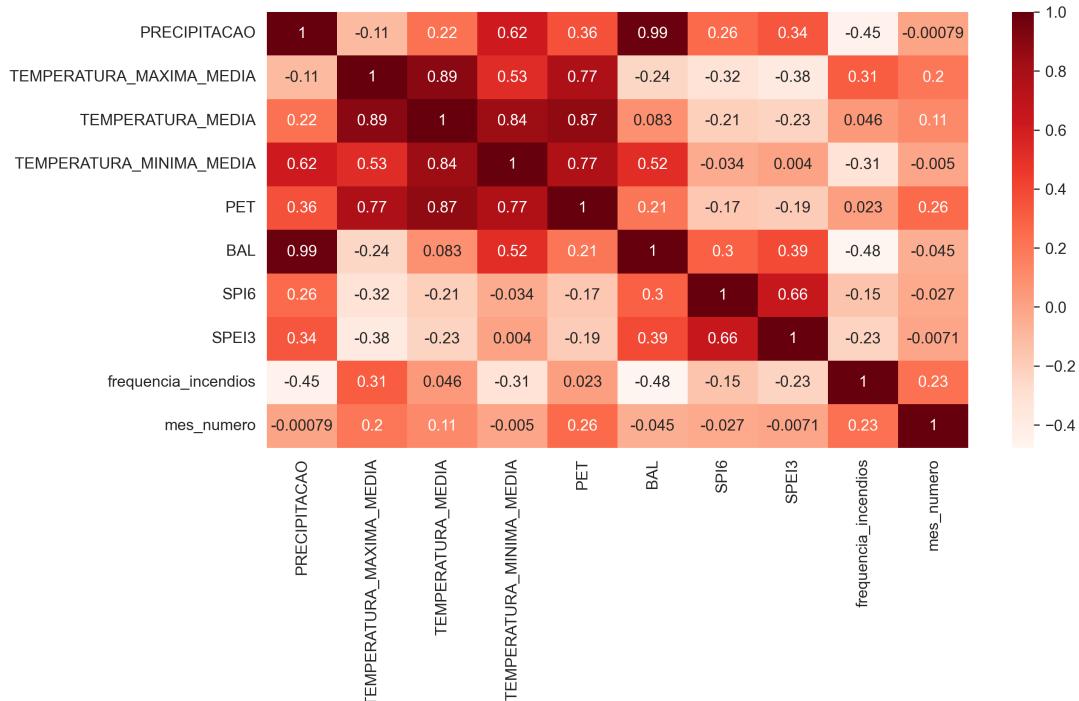


Figura 23 – Matriz de correlação

Logo, ao avaliar a correlação entre os pares de variáveis considerando a frequência de incêndios, decidiu-se pela escolha das seguintes variáveis para a implementação do modelo que irá prever a contabilização dos focos de incêndios: precipitação, SPEI3, temperatura máxima média e mês. É válido ressaltar a análise de multicolinearidade entre as variáveis, como, por exemplo, os índices SPI6 e SPEI3 do qual possuem um valor de correlação igual a 0.66, motivo do qual se decidiu pela escolha apenas do índice SPEI3. Por fim, torna-se necessária a tradução da variável mês para o formato categórico binário (dummy), uma vez que como identificado na análise exploratória, variáveis climáticas possuem sazonalidade.

6.6 Implementação dos modelos

A implementação do modelo é a etapa crucial após toda a análise elaborada da qual discorreu o texto anteriormente. Além disso, o objetivo desse estudo é alcançado nessa etapa, da qual ajuda a automatizar o processo de previsão dos quantitativos de foco dos incêndios por meio de variáveis climáticas, acarretando melhores estratégias e ações contra tais situações. Além disso, a implementação do modelo permite a realização de análises em tempo real, o que pode ser importante em tomadas de decisão por parte das autoridades responsáveis. No entanto, é importante garantir que o modelo seja implementado corretamente e validado antes de ser usado para tomar decisões críticas.

Dessa forma, a fim de verificar o melhor algoritmo para a implementação do modelo de previsão da quantidade de focos de incêndio por variáveis climáticas, das quais algumas foram previamente calculadas, fez-se necessário comparar diversos modelos estatísticos avaliando a melhor implementação por meio de métricas, pressupostos e testes com dados atuais. Consequentemente, decidiu-se comparar os seguintes modelos: regressão linear, random forest e arquiteturas de rede neural. Assim, será utilizado o R^2 a fim de verificar a explicabilidade do modelo criado e o RMSE (Raiz do Erro Médio Quadrático, em tradução livre) para identificar a acurácia do modelo sobre os dados, além dos respectivos pressupostos caso seja necessário. Portanto, decidiu-se avaliar o modelo considerando as variáveis escolhidas anteriormente, verificando combinações e as respectivas métricas R^2 e RMSE.

Vale ressaltar que, das métricas utilizadas, o R^2 (ou coeficiente de determinação) refere-se a quanto explicativo o modelo é em relação aos dados, e varia de 0 a 1, quanto mais próximo de 1, melhor as variáveis em questão conseguem explicar a variável alvo. Já o RMSE é uma das métricas para avaliar a acurácia das previsões através da raiz quadrada referente a diferença entre o treino e teste, porém a escolha por essa medida se deu pela interpretabilidade dos resultados, uma vez que o resultado pode ser interpretado como o desvio médio que as previsões têm do alvo.

Portanto, essa etapa visa demonstrar a implementação dos modelos citados ante-

riormente, isto é: regressão linear, random forest e rede neural. Vale ressaltar que todos os modelos foram realizados através da linguagem Python e as respectivas bibliotecas necessárias. Além disso, todos os modelos utilizaram as mesmas variáveis: precipitação, SPEI3, Temperatura Máxima Mensal e mês (variável dummy). Por fim, os modelos propostos utilizaram 80% dos dados para treino e 20% para teste, além de que houve a normalização pela média e desvio padrão (normal) nos dados, por conta das diferentes medidas numéricas para cada variável.

6.6.1 Regressão linear

Em relação ao modelo de regressão linear, fez-se a utilização do modelo múltiplo, uma vez que existem diversas variáveis explicativas. Dessa forma, as métricas do modelo gerado com as respectivas variáveis foram:

- **R²:** 0.64
- **RMSE:** 20.36

Além disso, verificou-se os pressupostos dos quais os modelos de regressão linear necessitam, como citado no referencial teórico. Portanto, apenas o pressuposto referente a homoscedasticidade falhou, sendo necessário, em primeiro momento, descartar o modelo linear proposto.

6.6.2 Random forest

Dessa forma, decidiu-se seguir para modelos não lineares, pois como evidenciado anteriormente, seriam necessários ajustes ou descarte nos dados, cenário do qual não será abordado. Vale ressaltar que, para encontrar os melhores valores para os hiperparâmetros desse modelo, isto é, os parâmetros cujos valores são utilizados para guiar o processo de aprendizado, utilizou-se a função GridSearchCV, do qual faz o ajuste dos hiperparâmetros utilizando Cross-Validation. Portanto, decidiu-se implementar o modelo random forest, do qual é um algoritmo que possui características não lineares, como citado no referencial teórico. Assim, as métricas do modelo criado com as respectivas variáveis estão apresentadas abaixo:

- **R²:** 0.71
- **RMSE:** 18.23

Como a literatura não apresenta pressupostos claros referente ao modelo em questão, decidiu-se por não verificar. Entretanto, percebe-se uma melhora significativa da métrica

R^2 ao implementar um modelo não linear, do qual as variáveis explicativas em questão conseguiram explicar em 71% a quantidade de incêndios florestais em Brasília-DF.

6.6.3 Rede neural

Por fim, definiu-se a arquitetura de rede neural como última abordagem referente a previsão de incêndios florestais por meio de variáveis climáticas. A arquitetura da rede neural utilizada possui três camadas densas com funções Relu de ativação, onde as duas primeiras apresentam regularização de Ridge e Lasso, respectivamente. Além disso, o otimizador de Adam foi utilizado, além da métrica do erro quadrático médio (MSE). Vale ressaltar que a rede em questão foi treinada em 100 épocas. Por fim, as métricas do modelo criado com as respectivas variáveis estão apresentadas abaixo:

- **R^2 :** 0.78
- **RMSE:** 17.19

É possível identificar que a rede neural possuiu a melhor explicabilidade entre os modelos propostos, do qual as variáveis independentes conseguem explicar 78% da variação da variável alvo, além de um erro de apenas 17.19 unidades. Por fim, ao testar o modelo com dados atuais referente ao ano de 2023, verificou-se um ótimo resultado, principalmente em meses com baixas frequências de incêndios, demonstrando a capacidade da rede neural criada.

6.6.4 Comparação dos modelos

A previsão de incêndios é uma tarefa crítica para a gestão e prevenção de desastres naturais, além do meio ambiente. Neste contexto, o texto propõe a utilização de modelos de aprendizado de máquina, como a regressão linear, floresta aleatória e rede neural para verificar a respectiva tarefa, dos quais foram implementados anteriormente. Neste sentido, comparar esses três modelos em relação à sua capacidade de prever a quantidade de incêndios com base em variáveis climáticas é uma questão relevante e pode fornecer insumos para a implementação de políticas públicas e, consequentemente, a definição de estudos ou estratégias a fim da prevenção dos focos de incêndios. Portanto, a fim de tornar a comparação mais visual, abaixo é apresentada uma tabela com os modelos e as respectivas métricas de comparação.

Portanto, ao verificar os modelos criados para as respectivas variáveis, isto é: Precipitação, SPEI3, Temperatura Máxima Média e mês (dummy), entende-se que o modelo de regressão linear múltipla deve ser descartado de imediato, uma vez que um dos pressupostos não foi atendido, além de apresentar as piores métricas entre os três. Da

Tabela 4 – Comparaçāo dos modelos

Modelos (Precipitação, SPEI3, Temperatura Máxima Média, Mês)	Pressupostos	RMSE	R ²
Regressão linear múltipla	Falhou em homoscedasticidade	20.36	0.64
Floresta aleatória	-	18.23	0.71
Rede neural	-	17.19	0.78

Fonte: autor, 2023

mesma forma, o modelo random forest apresentou interessantes métricas, motivo do qual pode ser explicado pelo fato dos dados inicialmente não demonstrarem comportamento linear. Assim, a rede neural com a respectiva arquitetura de três camadas foi o modelo com as melhores métricas, alcançando 78% de explicabilidade por meio das variáveis climáticas propostas. Portanto, os dois modelos últimos modelos devem ser aceitos para a implementação da previsão de incêndios florestais proposta no texto.

6.7 Aplicação do modelo selecionado em uma aplicação web

A fim de tornar a aplicação acessível e pública, elaborou-se uma simples aplicação web para a disponibilização dos resultados da pesquisa. Vale ressaltar que, a aplicação foi desenvolvida através do framework Streamlit, oriundo da linguagem de programação Python e hospedada na web através da própria nuvem do framework citado.

Dessa forma, a aplicação traz uma breve explicação sobre a necessidade da avaliação de queimadas no Distrito Federal com interesse de engajar o usuário. Além disso, é demonstrado gráficos dos quais abordam a variável alvo desse estudo: a quantidade de incêndios florestais, apresentando a série histórica e meses com maior frequência no período estudado.

À vista disso, no final da página é possível testar os modelos apresentados anteriormente nessa pesquisa, isto é: random forest e rede neural. Desse modo, torna-se possível ao usuário os ajustes nas variáveis em questão, das quais foram utilizadas para treinar o modelo: precipitação mensal prevista, temperatura máxima mensal prevista, índice SPEI3 e o mês desejado.

Consequentemente, ao aplicar os filtros desejados, os modelos imediatamente fazem os cálculos necessários e então demonstram os resultados referente a frequência de incêndios florestais de acordo com os valores escolhidos pelos usuários nos filtros disponíveis. Portanto, as figuras abaixo demonstram as telas referente à aplicação idealizada, a fim de engajar os usuários na pesquisa realizada.



Figura 24 – Introdução



Figura 25 – Visualização das séries históricas

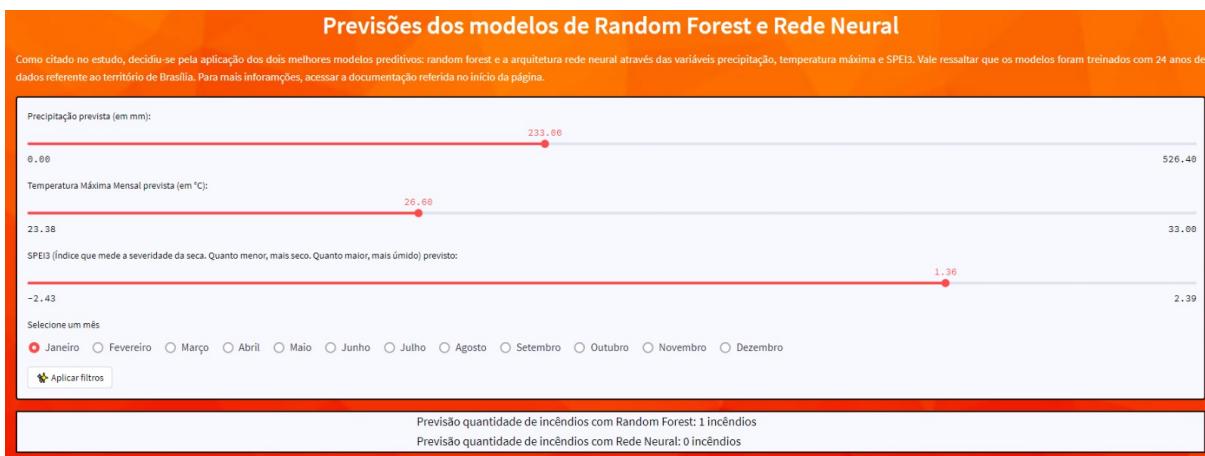


Figura 26 – Previsões

Então, a aplicação web elaborada nessa pesquisa visa o engajamento dos cidadãos com as análises realizadas, demonstrando como técnicas computacionais de aprendizado de máquina podem ser úteis ao meio ambiente. Vale ressaltar que, a aplicação pode ser acessada [clicando aqui](#). Da mesma forma, os códigos elaborados estão disponíveis no repositório GitHub do autor, do qual pode-se acessar [clicando aqui](#).

6.8 Resultados

Em relação aos incêndios ocorridos na capital durante os últimos 24 anos, identificaram-se anos dos quais obteram maior frequência em relação aos demais, isto é, os anos de 2007 e 2010. Em uma breve pesquisa diante as reportagens do passado, entendeu-se que o ano de 2007 na cidade de Brasília-DF obteve um dos recordes em relação à menor umidade do ar, alcançando a mínima de 10% ([BRAZILIENSE, 2023](#)). Já o ano de 2010 foi um dos anos dos quais a estiagem mais durou, alcançando 120 dias sem chuvas significativas na região ([PEDUZZI, 2023](#)).

Além disso, ao verificar os índices SPI6 e SPEI3 computados, é possível identificar a estiagem reportada em 2007 referente ao índice SPI6, do qual foi o ano em que o índice alcançou seu maior vale. Além disso, o índice SPI6 se sobressaiu em relação ao SPEI3, uma vez que deixou mais evidente a estiagem acumulada durante a série histórica proposta. Da mesma forma, o SPEI3 não demonstrou graficamente as estiagens da localidade de maneira clara, sendo necessário, possivelmente, ajustar a escala utilizada, uma vez que para tal utilizou-se três meses. Finalmente, o índice SPEI3 demonstrou uma melhor correlação em relação à quantidade de incêndios, possivelmente pela escala mais baixa.

Em seguida, o texto propôs a criação de modelos estatísticos: regressão linear, random forest e redes neurais, a fim de predizer a quantidade de incêndios florestais a partir das variáveis climáticas disponíveis na base de dados, da qual foi necessário a junção das tabelas em um período de intersecção entre as duas, isto é, de 1998 a 2022. Dessa forma, criou-se os três respectivos modelos considerando as variáveis: precipitação, temperatura máxima mensal e mês (dummy).

Ao avaliar os modelos, identificou-se que os pressupostos referentes a regressão linear não foram devidamente atendidos, além de que o modelo em questão foi o que apresentou piores resultados. Posteriormente, verificou-se a implementação do modelo random forest e redes neurais, dos quais alcançaram boas métricas de acurácia e erro relativo à previsão da quantidade de incêndios, decidindo descartar a regressão linear e seguir apenas com os modelos de random forest e redes neurais.

Consequentemente, ao testar o modelo com os dados atuais de 2023, verificaram-se bons resultados, principalmente quando considerada a sazonalidade da frequência de incêndios exposta na análise exploratória. Por exemplo, para o mês de março de 2023 houve uma precipitação de 98.23 mm, temperatura máxima mensal de 27.3°C e índice SPEI3 de 0.3, o modelo random forest previu apenas 1 incêndio, enquanto a rede neural fez a previsão de 0 incêndios, comportamento do qual também refletiu-se a acurácia das previsões nos meses do primeiro quartil de 2023.

Além disso, ao verificar a quantidade de incêndios de 2023 no portal do INPE referente ao mês de março, percebe-se que houveram 0 incêndios, demonstrando a qualidade

dos modelos em capturar a sazonalidade exposta. Por fim, decidiu-se pela implementação web para tornar os modelos implementados tangíveis e disponíveis a qualquer cidadão que se interesse pelo tema, do qual pode ser acessado [clicando aqui](#).

Portanto, entende-se que os resultados obtidos demonstram o alcance em relação ao objetivo do texto, a criação de um modelo estatístico para a previsão da quantidade de incêndios por variáveis climáticas e índices de secas no território de Brasília-DF.

7 Conclusões

No Brasil, cada vez mais se decidem medidas ambientais, principalmente referente a florestas e conservação das mesmas. Dessa forma, diversas políticas públicas e nacionais são articuladas, principalmente no Código Florestal brasileiro, a fim de criar estratégias e estudos para frear o avanço de incêndios nas florestas. Com isso, o texto conclui como a tecnologia e estatística podem ser aliadas ao combate dos focos de incêndios.

Consequentemente, o estudo obteve um escopo voltado apenas a localidade de Brasília-DF, cidade da qual possui o bioma cerrado, marcado pela longa estiagem durante o ano. Assim sendo, foi possível realizar análises minuciosas sobre as principais variáveis climáticas, disponibilizadas pelo INMET em um período de 30 anos. Ademais, tornou-se necessário a junção de tais dados com as contabilizações de queimadas, coletados através do INPE em um ciclo de 24 anos.

Além disso, buscou-se a possibilidade de implementação de índices de secas disponíveis na literatura, como Índice de Precipitação Padronizada (SPI) e Índice Padronizado de Precipitação-Evapotranspiração (SPEI), a fim de medir a severidade das secas dependendo da escala escolhida. Por meio desses, ao utilizar escalas de 6 e 3 meses respectivamente, identificou-se o comportamento da seca na localidade, correlacionando com impactos hídricos causados em situações de extrema seca durante o período de 30 anos.

Da mesma forma, com o intuito de alcançar o objetivo do estudo, decidiu-se pela implementação de modelos estatísticos dos quais fossem capazes de predizer a frequência de incêndios a partir das respectivas variáveis climáticas escolhidas, isto é: precipitação, temperatura máxima mensal e mês. À vista disso, o modelo de regressão foi descartado por não ter um de seus pressupostos atendidos, além das piores métricas de explicabilidade e erro. Portanto, os modelos estatísticos referentes a random forest e redes neurais apresentaram bons resultados nas respectivas métricas, e principalmente ao testá-lo com dados atuais referente ao primeiro quartil de 2023. Além disso, tornou-se público a interação com os modelos criados a partir de uma aplicação web, da qual pode ser acessado [clicando aqui](#).

Ademais, outras abordagens poderiam ser articuladas na continuidade desse estudo, como a avaliação de fenômenos climáticos, tal qual o El niñ, na análise de secas referente a localidade de Brasília-DF. Do mesmo modo, técnicas voltadas a séries temporais com regras de defasagem sobre os dados poderiam ser benéficas ao estudo proposto.

Portanto, conclui-se a possibilidade da implementação de modelos estatísticos para a previsão de incêndios florestais mensais mediante características climáticas, ferramenta da qual pode ser útil para as autoridades que desempenham estudos e ações relacionado ao tema em questão.

Referências

- ABRAMOWITZ, M.; STEGUN, I. A. Handbook of mathematical functions dover publications. *New York*, v. 361, 1965. Citado na página 37.
- AHMAD, M.; SINCLAIR, C.; WERRITTY, A. Log-logistic flood frequency analysis. *Journal of Hydrology*, Elsevier, v. 98, n. 3-4, p. 205–224, 1988. Citado na página 36.
- BRAZILIENSE, C. *Correio Braziliense Tempo: saiba quais foram os dias mais secos da história do DF*. 2023. Disponível em: <https://www.correobraziliense.com.br/cidades-df/2020/09/4873747-tempo-saiba-quais-foram-os-dias-mais-secos-da-historia-do-df.html>. Citado na página 63.
- BREIMAN, L. Random forests. *Machine learning*, Springer, v. 45, n. 1, p. 5–32, 2001. Citado na página 39.
- CHANG, L. H.; HUNSAKER, C. T.; DRAVES, J. D. Recent research on effects of climate change on water resources 1. *JAWRA Journal of the American Water Resources Association*, Wiley Online Library, v. 28, n. 2, p. 273–286, 1992. Citado na página 20.
- CHEIN, F. Introdução aos modelos de regressão linear: um passo inicial para compreensão da econometria como uma ferramenta de avaliação de políticas públicas. Escola Nacional de Administração Pública (Enap), 2019. Citado na página 38.
- FERNANDES, D. S. et al. Índices para a quantificação da seca. Santo Antônio de Goiás: Embrapa Arroz e Feijão, 2009., 2009. Citado 5 vezes nas páginas 28, 31, 32, 33 e 34.
- HOSKING, J. R. M.; WALLIS, J. R.; WOOD, E. F. Estimation of the generalized extreme-value distribution by the method of probability-weighted moments. *Technometrics*, Taylor & Francis, v. 27, n. 3, p. 251–261, 1985. Citado na página 36.
- JURAS, I. Aquecimento global e mudanças climáticas: uma introdução. *Plenarium*, v. 5, n. 5, p. 34–46, 2008. Citado na página 20.
- KEYANTASH, J. *NCAR Standardized Precipitation Index (SPI)*. 2023. Disponível em: <https://climatedataguide.ucar.edu/climate-data/standardized-precipitation-index-spi>. Citado na página 31.
- KEYANTASH, J.; DRACUP, J. A. The quantification of drought: an evaluation of drought indices. *Bulletin of the American Meteorological Society*, American Meteorological Society, v. 83, n. 8, p. 1167–1180, 2002. Citado na página 35.
- MCKEE, T. B. et al. The relationship of drought frequency and duration to time scales. In: BOSTON. *Proceedings of the 8th Conference on Applied Climatology*. [S.l.], 1993. v. 17, n. 22, p. 179–183. Citado 2 vezes nas páginas 28 e 34.
- MOLINA, P.; LIMA, L. Estudo de secas agrícolas no nordeste brasileiro. *Simpósio Brasileiro de Recursos Hídricos*, v. 13, 1999. Citado na página 28.
- OLIVEIRA, J. F. et al. Relação entre o standardized precipitation index (spi) e os relatórios de ocorrência de incêndios (roi) no parque nacional do itatiaia. *Floresta e Ambiente*, SciELO Brasil, v. 24, 2017. Citado na página 56.

PEDUZZI, S. C. e P. Agência Brasil Com 120 dias sem chuva, DF tem período mais longo de seca desde 2010. 2023. Disponível em: <<https://agenciabrasil.ebc.com.br/geral/noticia/2017-09/com-120-dias-sem-chuva-df-tem-periodo-mais-longo-de-seca-desde-2010>>. Citado na página 63.

SALVADOR, M. de A. Danos sociais e econômicos decorrentes de desastres naturais em consequência de fenômenos meteorológicos no brasil: 2010 – 2019. 2021. Citado na página 19.

SOUZA, L. Silva de et al. Air quality photochemical study over amazonia area, brazil. *International Journal of Environment and Pollution*, Inderscience Publishers, v. 48, n. 1-4, p. 194–202, 2012. Citado na página 19.

UPADHYAYA, L. N.; SINGH, H. P. Use of transformed auxiliary variable in estimating the finite population mean. *Biometrical Journal: Journal of Mathematical Methods in Biosciences*, Wiley Online Library, v. 41, n. 5, p. 627–636, 1999. Citado na página 36.

VICENTE-SERRANO, S. M. et al. A multiscalar drought index sensitive to global warming: the standardized precipitation evapotranspiration index. *Journal of climate*, American Meteorological Society, v. 23, n. 7, p. 1696–1718, 2010. Citado 4 vezes nas páginas 28, 35, 36 e 37.

Anexos

ANEXO A – Dicionário de Dados

A seguir estão descritos os dicionários das bases de dados utilizados.

A.1 Instituto Nacional de Pesquisas Espaciais (INPE) - Incêndios florestais

- **Ano:** Ano da respectiva contabilização mensal dos incêndios.
- **Mês:** Mês da respectiva contabilização mensal dos incêndios.
- **Frequência de incêndios:** A respectiva contabilização mensal dos incêndios.

A.2 Instituto Nacional de Meteorologia (INMET) - Estações Convencionais

- **Data Medição:** Data referente a coleta dos dados. Formato (YYYY-MM-DD).
- **Precipitação total (diário):** Total de precipitação (chuva) que ocorreu no espaço de 24 hora. Medida em milímetros (mm).
- **Temperatura Mínima (diário):** Mínima da temperatura do ar, ocorrida no período de 24 horas, a partir do termômetro específico para este fim. Medida em graus Celsius.
- **Temperatura Máxima (diário):** Máxima da temperatura do ar, ocorrida no período de 24 horas, a partir do termômetro específico para este fim. Medida em graus Celsius.
- **Temperatura Média (diário):** Média da temperatura do ar, ocorrida no período de 24 horas, a partir do termômetro específico para este fim. Medida em graus Celsius.

ANEXO B – Códigos dos Programas

A seguir estão todos os códigos utilizados para a criação desse trabalho.

B.1 Inserção e acesso dos dados no banco de dados - Python

Arquivo criado: InsercaoConsulta.ipynb Arquivos consumidos: inmet_1992_2023bsb.csv, historico_estado_distrito_federal.csv

- Arquivo criado: Insercao_consulta.ipynb
- Arquivos consumidos: inmet_1992_2023bsb.csv, historico_estado_distrito_federal.csv

Listing B.1 – Inserção e acesso dos dados no banco de dados - Python

```

1 # -*- coding: utf-8 -*-
2 """InsercaoAcessoDadosBD.ipynb
3
4 Automatically generated by Colaboratory.
5
6 Original file is located at
7     https://colab.research.google.com/drive/1Z23-qRtRtTAYqI46mirIbx5un7TC-rLP
8
9 # Inserção e acesso dados TCC 1
10 """
11
12 !pip install sqlalchemy
13 !pip install psycopg2
14
15 import pandas as pd
16 import numpy as np
17
18 import psycopg2
19 import sqlalchemy
20 from sqlalchemy import create_engine
21
22 """## Lendo os dados para inserir no PostgreSQL posteriormente"""
23
24 df = pd.read_csv("arquivo.csv")
25 df.head()
26

```

```
27 """## Inserindo dados no PostgreSQL"""
28
29 engine =
30     sqlalchemy.create_engine('postgresql://CONTA:SENHA@MAQUINA/NOME_BD_SERGIO_CRIOU')
31 con = engine.connect()
32
33 table_name = 'nome_tabela'
34 df.to_sql(table_name, con, if_exists = 'append', index = False)
35 #print(engine.table_names())
36
37 """## Acessando os dados do PostgreSQL"""
38
39 engine =
40     sqlalchemy.create_engine('postgresql://CONTA:SENHA@MAQUINA/NOME_BD_SERGIO_CRIOU')
41 df = pd.read_sql_query("SELECT * FROM nome_tabela", engine)
42 df.head()
```

B.2 Criação e análise índices SPI6 e SPEI3 - R

- Arquivo criado: SPI_SPEI.R
- Arquivos consumidos: inmet1992_2023bsb.csv
- Arquivos exportados: inmet_SPI_SPEI_indexes.csv

Listing B.2 – Criação e análise índices SPI6 e SPEI3 - R

```

1 install.packages("SPEI")
2 install.packages("readr")
3 library(SPEI)
4 library(readr)
5 library(glue)
6
7 options(scipen=999)
8
9 df <- read_delim("C:/Users/User/Desktop/inmet_teste3.csv",
10                      ";", escape_double = FALSE, locale =
11                      locale(decimal_mark = ","),
12                      trim_ws = TRUE)
13
14
15
16
17 ##### SPI #####
18 ##### SPI #####
19 ##### SPI #####
20
21
22 #SPI6 <- spi(df$'PRECIPITACAO',
23 #               scale = 6, na.rm = TRUE, dataopt = "monthly", cluster = TRUE,
24 #               gamma.dist = TRUE, start = c(1992, 1), end = c(2022, 12))
25
26 SPI6 <- spi(ts(df$'PRECIPITACAO', frequency = 12, start = c(1992, 1), end =
27               c(2022, 12)),
28               scale = 6, na.rm = TRUE, dataopt = "monthly", start = c(1992, 1),
29               end = c(2022, 12))
30
31 plot(SPI6)
32
33 ?spi

```

```

30
31 #Teste de normalidade
32 SPI_shapiro <- shapiro.test(SPI6$fitted)$p.value
33
34 #Histograma
35 hist(SPI6$fitted, xlab = 'Frequência', ylab = 'Índice SPI',
36       main = 'Frequência índice SPI-6 (Brasília - 1992 a 2022)',
37       col = 'cyan4',
38       breaks=c(-4, -2, -1.00, -0.5, 0.5, 1, 2, 4),
39       labels = TRUE)
40 lines(density(SPI6$fitted, na.rm = TRUE))
41 polygon(density(SPI6$fitted, na.rm = TRUE),
42           col=rgb(1,1,1,.2))
43 text(x = 0, y = 0.1, glue("Shapiro-Wilk p-valor: {round(SPI_shapiro,
44                           4)}"),cex=1)
45
46
47
48 ##### SPEI #####
49 ##### SPEI #####
50 #####
51 #df <- df [df$TEMPERATURA_MAXIMA_MEDIA <= 32,]
52 df$PET <- hargreaves(Tmin = df$TEMPERATURA_MINIMA_MEDIA,
53                       Tmax = df$TEMPERATURA_MAXIMA_MEDIA,
54                       lat = -15.789722)
55 df$BAL <- df$PRECIPITACAO - df$PET
56
57 #SPEI3 <- spei(df$BAL,
58 #               scale = 3, na.rm = TRUE, dataopt = "monthly", cluster = TRUE, start =
59 #               c(1992, 1), end = c(2022, 12))
60 SPEI3 <- spei(ts(df$BAL, frequency = 12, start = c(1992, 1), end = c(2022,
61                   12)),
62                   scale = 3, na.rm = FALSE, dataopt = "monthly", start = c(1992, 1), end
63                   = c(2022, 12))
64 plot(SPEI3)
65
66 #Teste de normalidade

```

```

67 SPEI_shapiro <- shapiro.test(SPEI3$fitted)$p.value
68
69 #Histograma
70 hist(SPEI3$fitted, xlab = 'Frequência', ylab = 'Índice SPEI',
71       main = 'Frequência índice SPEI-3 (Brasília - 1992 a 2022)',
72       col = 'cyan4', breaks=c(-4, -2, -1.00, -0.5, 0.5, 1, 2, 4),
73       labels=TRUE)
74 lines(density(SPEI3$fitted, na.rm = TRUE))
75 polygon(density(SPEI3$fitted, na.rm = TRUE),
76           col=rgb(1,1,1,.2))
77 text(x = 0, y = 0.1, glue("Shapiro-Wilk p-valor: {round(SPEI_shapiro,
78                           4)}"),cex=1)
79
80 df$SPI6 <- SPI6$fitted
81 df$SPEI3 <- SPEI3$fitted
82 write.csv(df, "inmet_SPI_SPEI_indexes.csv", row.names=FALSE)

```

B.3 Análise exploratória e modelagem - Python

- Arquivo criado: TCC Secas e Incêndios.ipynb
- Arquivos consumidos: inmet_1992_2023bsb.csv, inmet_SPI_SPEI_indexes.csv, historico_estado_distrito_federal.csv. Tabela temperaturas_incendios.
- Arquivos exportados: inmet_inpe.csv

Listing B.3 – Análise exploratória e modelagem - Python

```

1 # -*- coding: utf-8 -*-
2 """TCC Secas e Incêndios.ipynb
3
4 Automatically generated by Colaboratory.
5
6 Original file is located at
7     https://colab.research.google.com/drive/1s3vnUFW4VJCVfF07AaYbAt1saYzfP2MM
8
9 <div style="text-align:center;">
10    <a href="https://github.com/victoresende19" rel="some text">
11      

```

```
12 </a>
13
14 <h2 style="text-align:center">Análise de dados meteorológicos</h2>
15 <p style="text-align:center;font-size:13px;"></p>
16
17 <h2 style="text-align:center">Victor Augusto Souza Resende</h2>
18 <p style="text-align:center;font-size:13px;">Autor</p>
19 </div>
20
21 ## Sumário
22
23
24 - [1. Contexto] (#1)<br><br>
25
26 - [2. Análise exploratória INMET] (#2) <br>
27   * [2.1 Precipitação] (#21)<br>
28   * [2.2 Temperatura máxima] (#22)<br>
29   * [2.3 Temperatura mínima] (#23)<br>
30   * [2.4 Temperatura média] (#24)<br><br>
31
32 - [3. Análise exploratória INPE] (#2) <br>
33   * [3.1 Frequência incêndios] (#31) <br><br>
34
35 - [4. Manipulação tabelas INMET e INPE] (#4) <br>
36   * [4.1 Merge das tabelas] (#41)<br><br>
37
38 - [5. Modelagem] (#5) <br>
39   * [5.1 Padronização das variáveis] (#41)<br>
40   * [5.1 Random Forest] (#41)<br><br>
41 """
42
43 !pip install xgboost -q
44 !pip install tensorflow -q
45
46 import pandas as pd
47 import numpy as np
48
49 import matplotlib.pyplot as plt
50 from matplotlib.pyplot import figure
51 import seaborn as sns
52 from sklearn.tree import plot_tree
53
```

```
54 from scipy.stats import shapiro
55
56 from sklearn.model_selection import train_test_split
57 from sklearn.preprocessing import StandardScaler
58 from sklearn.metrics import mean_squared_error
59 from sklearn.metrics import mean_absolute_percentage_error
60 from sklearn.metrics import r2_score
61
62 #from sklearn.externals import joblib
63 import joblib
64
65 from sklearn.ensemble import RandomForestRegressor
66 from xgboost import XGBRegressor
67 from sklearn.ensemble import GradientBoostingRegressor
68 from sklearn.linear_model import LinearRegression, Lasso, Ridge
69 import tensorflow as tf
70
71 sns.set_style("whitegrid")
72
73 tf.__version__
74
75 """# 1 - Contexto <a id="1"></a>
76
77 Com o passar dos anos e o avanço da tecnologia, houve a criação e
    implementação de diversos sensores e sistemas em locais que correm riscos
    de desastres naturais para a verificação do perigo eminente e coleta de
    tais dados. Dessa forma, uma vasta quantidade de dados é gerada
    diariamente, principalmente quando há situações de estiagem, secas ou
    enchentes, dos quais são úteis para planos econômicos e sociais.
    Consequentemente, diversas técnicas de tratamento dos dados, modelagem e
    inferência estatística podem auxiliar na prevenção do impacto de
    determinado desastre natural caso ocorra.
78 """
79
80 inmet = pd.read_csv('../Data/inmet_teste3.csv', sep = ';', decimal = ',',
81                     parse_dates = ['Data Medicao'])
82 """# 2 - Análise exploratória INMET <a id="2"></a>"""
83
84 inmet.head()
85
86 inmet.describe()
```

```
87
88 inmet.info()
89
90 inmet.shape
91
92 round((inmet.isna().sum()/len(inmet)), 2)
93
94 """# 2.1 - Precipitação <a id="21"></a>"""
95
96 inmet[inmet['PRECIPITACAO'] == inmet['PRECIPITACAO'].max()]
97
98 len(inmet[inmet['PRECIPITACAO']] == inmet['PRECIPITACAO'].min())
99
100 figure(figsize=(18,5))
101 sns.lineplot(data=inmet, x="Data Medicao", y="PRECIPITACAO")
102 plt.xlabel('Data medição', fontsize=16)
103 plt.ylabel('Precipitação', fontsize=16)
104 plt.tick_params(axis='both', which='major', labelsize=18)
105 plt.savefig('../Documentação/precipitacaoTimeSeries.png', dpi=300)
106
107 figure(figsize=(10,5))
108 sns.histplot(data=inmet, x="PRECIPITACAO", kde=True)
109 plt.xlabel('Precipitação')
110 plt.ylabel('Frequência')
111 plt.savefig('../Documentação/precipitacaoHist.png', dpi=300,
    bbox_inches='tight')
112
113 import numpy as np
114 from scipy.stats import gamma, kstest
115
116 # Parmetros da distribuição gamma (estimados a partir dos dados de
117 # precipitação)
118 shape, loc, scale = gamma.fit(inmet['PRECIPITACAO'])
119
120 # Teste de Kolmogorov-Smirnov
121 p_valor, ks_statistic = kstest(inmet['PRECIPITACAO'], 'gamma', args=(shape,
    loc, scale))
122 print(f"Valor p: {p_valor:.4f}")
123
124 """# 2.2 - Temperatura máxima <a id="22"></a>"""
125
```

```
126 inmet[inmet['TEMPERATURA_MAXIMA_MEDIA']] ==
    inmet['TEMPERATURA_MAXIMA_MEDIA'].max()
127
128 inmet[inmet['TEMPERATURA_MAXIMA_MEDIA']] ==
    inmet['TEMPERATURA_MAXIMA_MEDIA'].min()
129
130 figure(figsize=(18,5))
131 sns.lineplot(data=inmet, x="Data Medicao", y="TEMPERATURA_MAXIMA_MEDIA")
132 plt.xlabel('Data medição', fontsize=16)
133 plt.ylabel('Temperatura máxima média', fontsize=16)
134 plt.tick_params(axis='both', which='major', labelsize=18)
135 plt.savefig('../Documentação/temperaturaMaximaTimeSeries.png', dpi=300)
136
137 figure(figsize=(10,5))
138 sns.histplot(data=inmet, x="TEMPERATURA_MAXIMA_MEDIA", kde=True)
139 plt.xlabel('Temperatura máxima média')
140 plt.ylabel('Frequência')
141 plt.savefig('../Documentação/temperaturaMaximaHist.png', dpi=300,
    bbox_inches='tight')
142
143 format(shapiro(inmet['TEMPERATURA_MAXIMA_MEDIA'].values)[1], '.8f')
144
145 """# 2.3 - Temperatura mínima <a id="23"></a>"""
146
147 inmet[inmet['TEMPERATURA_MINIMA_MEDIA']] ==
    inmet['TEMPERATURA_MINIMA_MEDIA'].max()
148
149 inmet[inmet['TEMPERATURA_MINIMA_MEDIA']] ==
    inmet['TEMPERATURA_MINIMA_MEDIA'].min()
150
151 figure(figsize=(18,5))
152 sns.lineplot(data=inmet, x="Data Medicao", y="TEMPERATURA_MINIMA_MEDIA")
153 plt.xlabel('Data medição', fontsize=16)
154 plt.ylabel('Temperatura mínima média', fontsize=16)
155 plt.tick_params(axis='both', which='major', labelsize=18)
156 plt.savefig('../Documentação/temperaturaMinimaTimeSeries.png', dpi=300)
157
158 figure(figsize=(10,5))
159 sns.histplot(data=inmet, x="TEMPERATURA_MINIMA_MEDIA", kde=True)
160 plt.xlabel('Temperatura mínima média')
161 plt.ylabel('Frequência')
162 plt.savefig('../Documentação/temperaturaMinimaHist.png', dpi=300,
```

```
    bbox_inches='tight')

163
164 format(shapiro(inmet['TEMPERATURA_MINIMA_MEDIA'].values)[1], '.18f')
165
166 from scipy.stats import normaltest
167
168 format(normaltest(inmet['TEMPERATURA_MINIMA_MEDIA'].values)[1], '.18f')
169
170 """# 2.4 - Temperatura média <a id="24"></a>"""
171
172 inmet[inmet['TEMPERATURA_MEDIA'] == inmet['TEMPERATURA_MEDIA'].max()]
173
174 inmet[inmet['TEMPERATURA_MEDIA'] == inmet['TEMPERATURA_MEDIA'].min()]
175
176 figure(figsize=(18,5))
177 sns.lineplot(data=inmet, x="Data Medicao", y="TEMPERATURA_MEDIA")
178 plt.xlabel('Data medição', fontsize=16)
179 plt.ylabel('Temperatura média', fontsize=16)
180 plt.tick_params(axis='both', which='major', labelsize=18)
181 plt.savefig('../Documentação/temperaturaMediaTimeSeries.png', dpi=300)
182
183 figure(figsize=(10,5))
184 sns.histplot(data=inmet, x="TEMPERATURA_MEDIA", kde=True)
185 plt.xlabel('Temperatura média')
186 plt.ylabel('Frequência')
187 plt.savefig('../Documentação/temperaturaMediaHist.png', dpi=300,
             bbox_inches='tight')
188
189 format(shapiro(inmet['TEMPERATURA_MEDIA'].values)[1], '.18f')
190
191 format(normaltest(inmet['TEMPERATURA_MEDIA'].values)[1], '.18f')
192
193 """# 2.5 - Correlação <a id="25"></a>"""
194
195 figure(figsize=(10,5))
196 sns.heatmap(inmet.corr(), annot=True)
197 plt.savefig('../Documentação/correlacao', dpi=300, bbox_inches='tight')
198
199 """# 2.6 - Outliers <a id="26"></a>"""
200
201 fig, axs = plt.subplots(ncols=4, figsize=(20,5))
```

```
203 # Plotagem dos boxplots
204 sns.boxplot(x=inmet['PRECIPITACAO'], ax=axs[0])
205 sns.boxplot(x=inmet['TEMPERATURA_MEDIA'], ax=axs[1])
206 sns.boxplot(x=inmet['TEMPERATURA_MINIMA_MEDIA'], ax=axs[2])
207 sns.boxplot(x=inmet['TEMPERATURA_MAXIMA_MEDIA'], ax=axs[3])
208 plt.savefig('../Documentação/boxplot', dpi=300, bbox_inches='tight')
209 plt.show()
210
211 """# 3 - Análise exploratória INPE <a id="3"></a>"""
212
213 # Carrega o arquivo CSV em um DataFrame
214 inpe = pd.read_csv('../Data/historico_estado_distrito_federal.csv', sep=',')
215 inpe = inpe.drop(columns=inpe.columns[-1])
216
217 # Transforma o DataFrame em um formato "tidy"
218 inpe = inpe.melt(id_vars=['Unnamed: 0'], var_name='mes',
                    value_name='frequencia')
219
220 # Renomeia a coluna "Unnamed: 0" para "ano"
221 inpe = inpe.rename(columns={'Unnamed: 0': 'ano'})
222
223 inpe.frequencia = inpe.frequencia.replace('-', 0).astype('int64')
224
225 # Exibe o DataFrame resultante
226 inpe.sort_values(by = ['ano']).reset_index(drop=True).head()
227
228 inpe.describe()
229
230 inpe.info()
231
232 inpe.shape
233
234 round((inpe.isna().sum()/len(inpe)), 2)
235
236 """# 3.1 - Frequência incêndios <a id="31"></a>"""
237
238 sns.lineplot(data=inpe, x="ano", y="frequencia", errorbar=None, markers=True,
                dashes=False)
239 plt.savefig('../Documentação/fireWild', dpi=300, bbox_inches='tight')
240 plt.show()
241
242 inpe[inpe.mes == 'Janeiro'].frequencia.mean()
```

```
243
244 figure(figsize=(20,5))
245 sns.lineplot(x = inpe.mes, y=inpe.frequencia, errorbar=None, markers=True,
246     dashes=False)
247 plt.savefig('.../Documentação/fireWildMonth', dpi=300, bbox_inches='tight')
248 plt.show()
249
250
251
252 plt.figure(figsize=(20, 5))
253 sns.lineplot(x=df_final.mes, y=df_final.PRECIPITACAO, errorbar=None,
254     markers=True, dashes=False, label='Precipitação')
255 sns.lineplot(x=inpe.mes, y=inpe.frequencia, errorbar=None, markers=True,
256     dashes=False, label='Quantidade de Incêndios')
257 plt.legend()
258 plt.savefig('.../Documentação/Precipitacao_and_fireWild_Month', dpi=300,
259     bbox_inches='tight')
260 plt.show()
261
262 """
263
264 inmet = pd.read_csv('..../Data/inmet_SPI_SPEI_indexes.csv')
265
266 # converte a coluna 'date' para o tipo datetime
267 inmet['Data Medicao'] = pd.to_datetime(inmet['Data Medicao'],
268                                         format='%m/%d/%Y')
269
270 # cria a nova coluna no formato desejado
271 inmet['Date'] = inmet['Data Medicao'].dt.strftime('%m/%Y')
272
273 inpe['mes_numero'] = inpe['mes'].replace({
274     'Janeiro': 1,
275     'Fevereiro': 2,
276     'Março': 3,
277     'Abril': 4,
278     'Maio': 5,
279     'Junho': 6,
     'Julho': 7,
```

```
280     'Agosto':8,
281     'Setembro':9,
282     'Outubro':10,
283     'Novembro':11,
284     'Dezembro':12
285 }, regex=True)
286
287 inpe['mes_numero'] = inpe['mes_numero'].astype(int)
288 inpe['ano'] = inpe['ano'].astype(str)
289
290 inpe['Date'] =
291     pd.to_datetime(inpe['mes_numero'].astype(str).str.cat(inpe["ano"], sep =
292         "/"), format='%m/%Y')
293 inpe['Date'] = inpe['Date'].dt.strftime('%m/%Y')
294
295 """# 4.1 - Merge das tabelas <a id="41"></a>"""
296
297 df_final = inmet.merge(inpe, on = 'Date')
298 #df_final = df_final.drop(columns=['PET', 'BAL'])
299 df_final = df_final.rename(columns={'frequencia':'frequencia_incendios'})
300 df_final.head()
301
302 df_final.to_csv('../Data/inmet_inpe.csv')
303
304 figure(figsize=(20,5))
305 sns.lineplot(x = df_final.mes, y=df_final.PRECIPITACAO, errorbar=None,
306                 markers=True, dashes=False)
307 plt.savefig('../Documentação/PrecipitacaoMonth', dpi=300,
308             bbox_inches='tight')
309 plt.show()
310
311 """# 5 - Modelagem <a id="5"></a>"""
312
313 figure(figsize=(10,5))
314 sns.heatmap(df_final.corr(), annot = True, cmap="Reds")
315 plt.savefig('../Documentação/CorrelacaoFinal', dpi=300, bbox_inches='tight')
316 plt.show()
317
318 # lags = 2 # número de atrasos
319 # for i in range(1, lags+1):
320 #     df_final[f'lag_{i}'] = df_final['frequencia_incendios'].shift(i)
```

```
318 # df_final.dropna(inplace=True)
319
320 df_final = pd.get_dummies(df_final, columns = ['mes'])
321
322 cols = ['PRECIPITACAO',
323          'SPEI3',
324          'TEMPERATURA_MAXIMA_MEDIA',
325          'mes_Janeiro',
326          'mes_Fevereiro',
327          'mes_Março',
328          'mes_Abril',
329          'mes_Maio',
330          'mes_Junho',
331          'mes_Julho',
332          'mes_Agosto',
333          'mes_Setembro',
334          'mes_Outubro',
335          'mes_Novembro',
336          'mes_Dezembro']
337 X = df_final.loc[:, cols].values
338 y = df_final.loc[:, 'frequencia_incendios'].values
339
340 X_train, X_test, y_train, y_test = train_test_split(X, y, test_size = 0.2,
341                                                 random_state = 0)
342 """# 5.1 - Padronização das variáveis <a id="51"></a>"""
343
344 ss = StandardScaler()
345 X_stand_train = ss.fit_transform(X_train)
346 X_stand_test = ss.transform(X_test)
347 #joblib.dump(ss, '../Models/padronizacao')
348
349 """# 5.2 - Linear Regression <a id="52"></a>"""
350
351 linear = LinearRegression()
352 linear.fit(X_stand_train, y_train)
353
354 y_pred = linear.predict(X_stand_test)
355 score_stand_linear = linear.score(X_stand_test, y_test)
356 mse_linear = mean_squared_error(y_test, y_pred)
357 rmse_linear = mean_squared_error(y_test, y_pred, squared=False)
358
```

```
359 print(f'R2 Linear Regression: {score_stand_linear:.2f}')
360 print(f'MSE: {mse_linear:.2f}')
361 print(f'RMSE: {rmse_linear:.2f}')
362
363 value_dezembro = np.array([233.0, 1.357870, 26.60, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0]).reshape(-1, 15)
364 value_agosto = np.array([0.0, -0.672021, 27.7, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0]).reshape(-1, 15)
365
366 print(f'Previsão quantidade incêndios agosto:
367     {linear.predict(ss.transform(value_agosto))[0]:.0f}')
367 print(f'Previsão quantidade incêndios dezembro:
368     {linear.predict(ss.transform(value_dezembro))[0]:.0f}')
369 """# 5.3 - Random Forest <a id="53"></a>"""
370
371 # import pandas as pd
372 # from sklearn.model_selection import train_test_split
373 # from sklearn.ensemble import RandomForestRegressor
374 # from sklearn.metrics import mean_squared_error
375 # from sklearn.preprocessing import StandardScaler
376
377 # # Adicionando uma coluna com a frequência de incêndios defasada em um mês
378 # df_final['frequencia_incendios_lag'] =
379 #     df_final['frequencia_incendios'].shift(1)
380 # df_final.dropna(inplace=True)
381
382 # # Separando os dados em treino e teste
383 # cols = ['PRECIPITACAO',
384 #         'SPEI3',
385 #         'TEMPERATURA_MAXIMA_MEDIA',
386 #         'mes_Janeiro',
387 #         'mes_Fevereiro',
388 #         'mes_Março',
389 #         'mes_Abril',
390 #         'mes_Maio',
391 #         'mes_Junho',
392 #         'mes_Julho',
393 #         'mes_Agosto',
394 #         'mes_Setembro',
395 #         'mes_Outubro',
395 #         'mes_Novembro',
```

```
396 #         'mes_Dezembro',
397 #         'frequencia_incendios_lag']
398
399 # X = df_final.loc[:, cols].values
400 # y = df_final.loc[:, 'frequencia_incendios'].values
401
402 # X_train, X_test, y_train, y_test = train_test_split(X, y, test_size = 0.2,
403 #                                                       random_state = 0)
404
405 # # Normalizando os dados
406 # ss = StandardScaler()
407 # X_stand_train = ss.fit_transform(X_train)
408 # X_stand_test = ss.transform(X_test)
409
410 # # Criando o modelo de Random Forest Regressor
411 # regressor = RandomForestRegressor(bootstrap= True,
412 #                                     max_depth= 50,
413 #                                     max_features= 2,
414 #                                     min_samples_leaf= 2,
415 #                                     min_samples_split= 5,
416 #                                     n_estimators= 200)
417
418 # # Treinando o modelo
419 # regressor.fit(X_stand_train, y_train)
420
421 # # Fazendo as previsões
422 # y_pred = regressor.predict(X_stand_test)
423
424 # # Calculando as métricas
425 # score_stand_ran = regressor.score(X_stand_test, y_test)
426 # mse = mean_squared_error(y_test, y_pred)
427 # rmse = mean_squared_error(y_test, y_pred, squared = False)
428
429 # # Imprimindo as métricas
430 # print(f'R2 Random Forest Regressor: {score_stand_ran:.2f}')
431 # print(f'MSE: {mse:.2f}')
432 # print(f'RMSE: {rmse:.2f}')
433
434 # # Dados de abril (valores conhecidos)
435 # X_april = np.array([10, 0.5, 25, 0, 0, 0, 1, 0, 0, 0, 0, 0, 0, 0, 0,
436 #                     1]).reshape(1, -1)
```

```
436 # # Dados de maio (valores desconhecidos)
437 # X_may = np.array([np.nan, np.nan, np.nan, 0, 0, 0, 0, 1, 0, 0, 0, 0, 0,
438 #   0, 1]).reshape(1, -1)
439
440 # # Imputação dos valores de maio usando a média dos valores de abril
441 # mean_april = np.mean(X_april[:, :3], axis=0)
442 # mean_april_tiled = np.tile(mean_april, (X_may.shape[0], 1))
443 # X_may[:, :3] = np.nan_to_num(X_may[:, :3], nan=mean_april_tiled)
444
445 # # Padronização dos dados
446 # X_stand_april = ss.transform(X_april)
447 # X_stand_may = ss.transform(X_may)
448
449 # # Previsão da quantidade de incêndios em maio
450 # y_pred_may = regressor.predict(X_stand_may)
451
452 # print(f'Previsão da quantidade de incêndios em maio: {y_pred_may[0]:.2f}')
453
454 regressor = RandomForestRegressor(bootstrap= True,
455                                     max_depth= 50,
456                                     max_features= 2,
457                                     min_samples_leaf= 2,
458                                     min_samples_split= 5,
459                                     n_estimators= 200)
460
461 regressor.fit(X_stand_train, y_train)
462 y_pred = regressor.predict(X_stand_test)
463
464 score_stand_ran = regressor.score(X_stand_test, y_test)
465 mse = mean_squared_error(y_test, y_pred)
466 rmse = mean_squared_error(y_test, y_pred, squared = False)
467
468 #Métricas
469 print(f'R2 Random Forest Regressor: {score_stand_ran:.2f}')
470 print(f'MSE: {mse:.2f}')
471 print(f'RMSE: {rmse:.2f}')
472
473 #Exporta e importa o modelo
474 #joblib.dump(regressor, '../Models/floresta_aleatoria_shift')
475 #regressor = joblib.load('../Models/floresta_aleatoria_shift')
476
477 value_dezembro = np.array([233.0, 1.357870, 26.60, 0, 0, 0, 0, 0, 0, 0, 0, 0,
```

```
    0, 0, 1]).reshape(-1, 15)
477 value_agosto = np.array([0.0, -0.672021, 27.7, 0, 0, 0, 0, 0, 1, 0, 0, 0, 0, 0,
    0, 0]).reshape(-1, 15)
478
479 print(f'Previsão quantidade incêndios agosto:
        {regressor.predict(ss.transform(value_agosto))[0]:.0f}')
480 print(f'Previsão quantidade incêndios dezembro:
        {regressor.predict(ss.transform(value_dezembro))[0]:.0f}')
481
482 # n_arvore = 0
483
484 # fig, axes = plt.subplots(nrows = 1,ncols = 1,figsize = (15,15), dpi=800)
485 # plot_tree(regressor.estimators_[n_arvore],
486 #             feature_names=['PRECIPITACAO', 'SPEI3',
487 #                             'TEMPERATURA_MAXIMA_MEDIA', 'TEMPERATURA_MINIMA_MEDIA', 'BAL',
488 #                             'mes_numero'],
489 #             class_names=['frequencia_incendios'],
490 #             filled=True, impurity=True, proportion = True, fontsize = 5,
491 #             rounded=False, max_depth = 3)
492 # plt.savefig(f'../Documentação/Modelo-Arvore-LongProf-Numero-{n_arvore}.png',
493 #             dpi=300, bbox_inches='tight')
494
495 # fig, axes = plt.subplots(nrows = 1,ncols = 5,figsize = (40,25), dpi=900)
496 # for i in range(0, 5):
497 #     tree.plot_tree(
498 #         regressor.estimators_[i],
499 #         feature_names=['PRECIPITACAO', 'SPEI3', 'TEMPERATURA_MAXIMA_MEDIA',
500 #                         'TEMPERATURA_MINIMA_MEDIA', 'BAL', 'mes_numero'],
501 #         class_names=['frequencia_incendios'],
502 #         filled=True, impurity=True, proportion = True, fontsize = 5,
503 #         rounded=False, max_depth = 3, ax = axes[i])
504
505 #     axes[index].set_title('Estimator: ' + str(i), fontsize = 20)
506 # fig.savefig('Modelo-Arvores-0-2.png')
507
508 """# 5.4 - Rede Neural <a id="54"></a>"""
509
510 # Definindo o modelo de RNA com regularização L1 e L2
511 model = tf.keras.Sequential([
512     tf.keras.layers.Dense(64, activation='relu', input_shape=(X_train.shape[1],),
513     kernel_regularizer=tf.keras.regularizers.l1_l2(l1=0.01, l2=0.01)),
```

```
510     tf.keras.layers.Dense(32, activation='relu',
511
512         kernel_regularizer=tf.keras.regularizers.l1_l2(l1=0.01, l2=0.01)),
513     tf.keras.layers.Dense(1, activation='relu')
514 )
515
516 model.compile(optimizer='adam', loss='mse', metrics=['mse', 'mape'])
517
518 # Treinando o modelo
519 history = model.fit(X_stand_train, y_train, epochs=100,
520                       validation_data=(X_stand_test, y_test))
521
522 # Fazendo previsões com o modelo treinado
523 y_pred = model.predict(X_stand_test)
524
525 # Calculando as métricas de desempenho
526 score_stand_nn = r2_score(y_test, y_pred)
527 mse_nn = mean_squared_error(y_test, y_pred)
528 rmse_nn = mean_squared_error(y_test, y_pred, squared=False)
529
530 #Exporta e importa o modelo
531 #joblib.dump(model, '../Models/rede_neural1')
532 #model = joblib.load('../Models/rede_neural1')
533
534 #import pickle
535 #pickle.dump(model, open('../Models/rede_neural1.pkl', 'wb'))
536 #model = pickle.load(open('../Models/rede_neural1.pkl', 'rb'))
537
538 #from tensorflow.keras.models import load_model
539 #model.save('redes_neurais_shift.h5')
540 #model = load_model('redes_neurais_shift.h5')
541
542 value_dezembro = np.array([233.0, 1.357870, 26.60, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 1]).reshape(-1, 15)
543 value_agosto = np.array([0.0, -0.672021, 27.7, 0, 0, 0, 0, 0, 0, 0, 0, 1, 0, 0, 0, 0, 0]).reshape(-1, 15)
544 value_maio = np.array([144.0, 0.893, 27.15, 0, 0, 0, 0, 1, 0, 0, 0, 0, 0, 0, 0, 0, 0]).reshape(-1, 15)
545
546 print(f'R2 Neural Network Regression: {score_stand_nn:.2f}')
```

```
547 print(f'MSE: {mse_nn:.2f}')
548 print(f'RMSE: {rmse_nn:.2f}')
549 print(f'Previsão quantidade incêndios dezembro:
      {model.predict(ss.transform(value_dezembro))[0][0]:.0f}')
550 print(f'Previsão quantidade incêndios agosto:
      {model.predict(ss.transform(value_agosto))[0][0]:.0f}')
551 print(f'Previsão quantidade incêndios maio:
      {model.predict(ss.transform(value_maio))[0][0]:.0f}')
552
553 """# 6 - Séries temporais <a id="6"></a>"""
554
555 # Importar bibliotecas
556 import pandas as pd
557 import numpy as np
558 import matplotlib.pyplot as plt
559 from statsmodels.tsa.statespace.sarimax import SARIMAX
560 from statsmodels.graphics.tsaplots import plot_acf, plot_pacf
561 from statsmodels.tsa.stattools import adfuller
562
563 # Importar dados
564 dados = pd.read_csv('inmet_inpe.csv', index_col='Data Medicao',
      parse_dates=True)
565 dados = dados['frequencia_incendios']
566
567 # Verificar estacionariedade
568 result = adfuller(dados)
569 print('ADF Statistic: %f' % result[0])
570 print('p-value: %f' % result[1])
571 if result[1] > 0.05:
572     print('Série não é estacionária')
573 else:
574     print('Série é estacionária')
575
576 # Gráficos de autocorrelação e autocorrelação parcial
577 plot_acf(dados)
578 plot_pacf(dados)
579 plt.show()
580
581 # Ajustar modelo SARIMA
582 modelo = SARIMAX(dados, order=(2, 0, 0), seasonal_order=(2, 0, 0, 6))
583 resultado = modelo.fit()
```

```
585 # Fazer previsões
586 previsoes = resultado.predict(start=len(dados), end=len(dados)+11,
587                               dynamic=False)
588 # Plotar previsões
589 plt.plot(dados.index, dados, label='Dados')
590 plt.plot(previsoes.index, previsoes, label='Previsões')
591 plt.legend()
592 plt.show()
593
594 """<hr>
595
596 Victor Resende
597 """
```