UNIVERSIDADE FEDERAL DE JUIZ DE FORA
INSTITUTO DE CIÊNCIAS EXATAS
PROGRAMA DE PÓS-GRADUAÇÃO EM MODELAGEM COMPUTACIONAL

Túlio José Francisco

Modelagem de Secas Usando um Modelo de Expansão Polinomial Evolutiva Interpretável com Seleção de Características

Juiz de Fora
2024

Túlio José Francisco

Modelagem de Secas Usando um Modelo de Expansão Polinomial Evolutiva Interpretável com Seleção de Características

Dissertação apresentada ao PROGRAMA DE PÓS-GRADUAÇÃO EM MODELAGEM COMPUTACIONAL da Universidade Federal de Juiz de Fora como requisito parcial à obtenção do título de Mestre em Modelagem Computacional Modelagem Computacional:

Orientador: Doutor Leonardo Goliatt da Fonseca

Coorientador: Doutora Camila Martins Saporetti

Juiz de Fora

2024

Ficha catalográfica elaborada através do Modelo Latex do CDC da UFJF
com os dados fornecidos pelo(a) autor(a)

Francisco, Túlio José.
Modelagem de Secas Usando um Modelo de Expansão Polinomial Evolutiva Interpretável com Seleção de Características / Túlio José Francisco. – 2024.
63 f. : il.

Orientador: Leonardo Goliatt da Fonseca
Coorientador: Camila Martins Saporetti
Dissertação (Mestrado) – Universidade Federal de Juiz de Fora, INSTITUTO DE CIÊNCIAS EXATAS. PROGRAMA DE PÓS-GRADUAÇÃO EM MODELAGEM COMPUTACIONAL, 2024.

1. Secas. 2. Modelagem. 3. Expansão Polinomial. I. Goliatt, Leonardo, orient. II. Saporetti, Camila , coorient. Título.

Túlio José Francisco

Modelagem de secas usando um modelo de expansão polinomial evolutiva interpretável com seleção de características

Dissertação apresentada ao Programa de Pós-Graduação em Modelagem Computacional da Universidade Federal de Juiz de Fora como requisito parcial à obtenção do título de Mestre em Modelagem Computacional. Área de concentração: Modelagem Computacional.

Aprovada em 09 de setembro de 2024.

BANCA EXAMINADORA

Prof. Dr. Leonardo Goliatt da Fonseca - Orientador
Universidade Federal de Juiz de Fora

Prof.ª Drª. Camila Martins Saporetti - Coorientadora
Universidade do Estado do Rio de Janeiro

Prof. Dr. Wanderlei Malaquias Pereira Júnior
Universidade Federal de Catalão

Prof.ª Drª. Eliane da Silva Christo
Universidade Federal de Juiz de Fora

Juiz de Fora, 21/08/2024.

Documento assinado eletronicamente por **Leonardo Goliatt da Fonseca**, **Professor(a)**, em 09/09/2024, às 11:29, conforme horário oficial de Brasília, com fundamento no § 3º do art. 4º do Decreto nº 10.543, de 13 de novembro de 2020.

Documento assinado eletronicamente por **Camila Martins Saporetti**, **Usuário Externo**, em 09/09/2024, às 11:32, conforme horário oficial de Brasília, com fundamento no § 3º do art. 4º do Decreto nº 10.543, de 13 de novembro de 2020.

Documento assinado eletronicamente por **Wanderlei Malaquias Pereira Junior**, **Usuário Externo**, em 09/09/2024, às 11:34, conforme horário oficial de Brasília, com fundamento no § 3º do art. 4º do Decreto nº 10.543, de 13 de novembro de 2020.

Documento assinado eletronicamente por **Eliane da Silva Christo**, **Professor(a)**, em 09/09/2024, às 13:29, conforme horário oficial de Brasília, com fundamento no § 3º do art. 4º do Decreto nº 10.543, de 13 de novembro de 2020.

A autenticidade deste documento pode ser conferida no Portal do SEI-Ufjf (www2.ufjf.br/SEI) através do ícone Conferência de Documentos, informando o código verificador **1938226** e o código CRC **283A5E3C**.

AGRADECIMENTOS

Ao concluir este trabalho, gostaria de expressar minha profunda gratidão a todos que, de alguma forma, contribuíram para a realização desta dissertação.

Primeiramente, agradeço a Deus, por me dar força e sabedoria ao longo dessa jornada acadêmica.

Aos meus orientadores, Leonardo Goliatt e Camila Saporetti, pela orientação dedicada, paciência, e pelo constante incentivo durante todo o processo de pesquisa e redação. Suas valiosas sugestões e conselhos foram essenciais para o desenvolvimento deste trabalho.

Aos professores do Programa de Pós-Graduação em Modelagem Computacional da Universidade Federal de Juiz de Fora, por compartilharem seus conhecimentos e por contribuírem significativamente para minha formação acadêmica. Principalmente a coordenadora do programa, Flávia Bastos, por todo o suporte dado para que esse trabalho se tornasse realidade.

À minha família, especialmente aos meus pais, Turanbem e Erenilda e meu irmão Turanbem Jr., pelo amor incondicional, apoio emocional, e por acreditarem em mim, mesmo nos momentos de dúvida. Sem vocês, nada disso teria sido possível.

Aos meus amigos, Bruno, Lucas e Luis pelas conversas, risadas e por serem uma fonte constante de inspiração e apoio durante essa jornada acadêmica.

Por fim, a todos que, direta ou indiretamente, contribuíram para a concretização deste trabalho, deixo aqui meu sincero agradecimento.

RESUMO

As secas são fenômenos naturais capazes de gerar consequências negativas tanto para o meio ambiente quanto para a sociedade humana, como escassez de água, perda de colheitas, incêndios florestais e até desertificação do solo. A literatura fornece vários índices para monitoramento de secas, como o Índice de Precipitação Padronizada (SPI), que visa determinar períodos secos e úmidos em escalas de tempo que variam de 3 a 48 meses. Esta dissertação apresenta uma abordagem para a modelagem de secas ao introduzir um modelo de Expansão Polinomial Evolutiva (EPE) com técnicas integradas de seleção de características. Utilizando o EPE é possível gerar um modelo de previsão com explicabilidade. O modelo EPE, uma importante ferramenta matemática, é empregado para aumentar a precisão da previsão de secas. O EPE foi avaliado em um conjunto de dados de seca da Turquia. O EPE pode capturar a dinâmica complexa da seca e identificar as características mais importantes para a previsão de secas, ou seja, é uma ferramenta valiosa para a previsão e mitigação de secas. Esta informação pode ser usada para desenvolver sistemas de alerta precoce para secas e estratégias para mitigar os impactos da seca. Nesse trabalho o EPE foi aplicado em conjunto com a regressão Lasso para criar polinômios interpretáveis a partir dos dados de entrada, assim gerando um modelo de previsão de secas com explicabilidade e métricas satisfatórias.

Palavras chave: Secas; Modelagem; Expansão Polinomial, Explicabilidade

ABSTRACT

Droughts are natural phenomena capable of generating negative consequences for both the environment and human society, such as water scarcity, crop loss, forest fires, and even soil desertification. The literature provides various indices for monitoring droughts, such as the Standardized Precipitation Index (SPI), which aims to determine dry and wet periods on time scales ranging from 3 to 48 months. This paper presents a novel approach to drought modeling by introducing an Evolutionary Polynomial Expansion (EPE) model with integrated feature selection techniques. The EPE model, a powerful mathematical tool, is employed to enhance the accuracy of drought prediction. The EPE was evaluated on a dataset of drought data from Turkey. It can potentially be a valuable tool for drought forecasting and mitigation. The EPE can capture the complex dynamics of drought and identify the most important features for drought prediction. This information can be used to develop early warning systems for drought and strategies for mitigating drought impacts.

Keywords: Drought; EPE; Explainable AI.

LISTA DE ILUSTRAÇÕES

LISTA DE TABELAS

LISTA DE ABREVIATURAS E SIGLAS

SPI	*Standard Precipitation Index*
EDI	*Effective Drought Index*
XAI	*Explainable Artificial Intelligence*
EPE	*Evolutionary Polynomial Expansion*
MAE	*Mean Absolute Error*
MAPE	*Mean Absolute Percentage Error*
MSE	*Mean Squared Error*
NSE	*Nash-Sutcliffe Efficiency*
RMSE	*Root Mean Squared Error*
RRMSE	*Relative Root Mean Squared Error*
R^2	*Coefficient of Determination*
WI	*Willmott Index*
KGE	*Kling-Gupta Efficiency*
ACF	*Autocorrelation Function*
PACF	*Partial Autocorrelation Function*
ARIMA	*AutoRegressive Integrated Moving Average*
RNN	*Recurrent Neural Network*
LSTM	*Long Short-Term Memory*
CNN	*Convolutional Neural Network*
GP	*Genetic Programming*
PYSR	*Python Symbolic Regression*

LISTA DE SÍMBOLOS

$\forall$	Para todo
$\in$	Pertence
σ	Desvio padrão
X	Média histórica da precipitação
γ	Proporção entre as médias dos valores simulados e observados
α	Proporção entre os desvios padrão dos valores simulados e reais
β	Parâmetro de controle de complexidade das expressões polinomiais

SUMÁRIO

1 INTRODUÇÃO

A seca está entre as calamidades climáticas mais destrutivas que ameaçam ecossistemas e a sociedade, devido às graves consequências para a quantidade e qualidade da água em escalas menores e maiores. A seca surge de uma escassez de precipitação, e outros fatores, como o aumento das temperaturas e o crescimento do consumo humano de água, que podem agravar esse evento (6). O déficit de precipitação pode se acumular rapidamente em um curto período de tempo ou levar meses para se manifestar através de fluxos de rios reduzidos, níveis de reservatórios mais altos ou lençóis freáticos subterrâneos mais profundos.

Para estimar a gravidade da seca, os indicadores relativos são medidos empregando um conjunto de informações sobre meteorologia, agricultura, hidrologia ou socioeconomia, além de fatores relacionados ao clima, como temperatura e precipitação (7). Esses índices avaliam a severidade da seca em desenvolvimento e, se apoiados pela estrutura institucional e funções adequadas, podem ser utilizados para acionar planos de contingência contra a seca.

Alguns indicadores de seca baseados em precipitação foram propostos, como o índice padronizado de precipitação (SPI) (5) e o índice efetivo de seca (EDI) (8). No entanto, embora sejam funcionais para determinar sinais prévios de secas, eles apenas identificam casos já ocorridos. O desafio é estimar os dias futuros de seca, semanas, meses, e sua intensidade, ou seja, aumentar o potencial de alerta dos sistemas de monitoramento de seca por meio da previsão de seca. Embora esses índices sejam inicialmente meteorológicos, sua medida seria instrutiva na administração da seca agrícola e no desempenho dos recursos hídricos.

Várias técnicas e abordagens computacionais para previsão de seca foram propostas e aplicadas em zonas distintas nos últimos anos. Devido à importância da seca na vida da população, várias pesquisas recentes foram realizadas para aumentar a precisão dos métodos de previsão de seca (9). Nesse contexto, o potencial de modelos de aprendizado de máquina (ML) distintos, como Árvore de Decisão, Gradient Boosting, Floresta Aleatória e Máquina de Vetor de Suporte, entre outros, foi substancialmente analisado.

1.1 MOTIVAÇÃO

A previsão de secas é uma tarefa complexa e desafiadora, dado o caráter multifacetado e não linear dos fatores que contribuem para a ocorrência e severidade desses eventos. A variabilidade climática, associada a fatores como precipitação, temperatura, e umidade, torna difícil a aplicação de métodos tradicionais de modelagem que dependem de suposições lineares e simplificações dos processos naturais. Como resultado, os índices de seca comumente utilizados, como o Índice Padronizado de Precipitação (SPI), ainda

que eficazes na identificação de secas passadas, apresentam limitações quando aplicados à previsão de eventos futuros, especialmente em escalas temporais mais longas (5). Esse contexto exige o desenvolvimento de abordagens mais robustas e adaptativas, capazes de capturar a complexidade e a dinâmica intrínseca dos sistemas climáticos.

Nesse cenário, técnicas de aprendizado de máquina (ML) surgem como uma solução promissora para melhorar a acurácia das previsões de seca. Modelos como árvores de decisão, florestas aleatórias, e máquinas de vetor de suporte têm sido amplamente estudados e aplicados para esse fim, proporcionando uma capacidade superior de identificar padrões ocultos nos dados e de lidar com a natureza não linear dos processos envolvidos (10). O modelo de Expansão Polinomial Evolutiva (EPE) se destaca nesse contexto, pois combina a flexibilidade e o poder de predição do aprendizado de máquina com a interpretabilidade das expressões polinomiais. Ao evoluir polinômios de forma iterativa e seletiva, polinômios construídos evolutivamente permitem a construção de modelos que não apenas são precisos, mas também oferecem uma compreensão clara das relações subjacentes entre as variáveis como mostrado por (11) no contexto de modelagem de escoamento sob barragens de aço.

1.2 OBJETIVOS

Este estudo visa avançar o campo da modelagem de secas através do desenvolvimento e validação de abordagens híbridas de aprendizado de máquina que priorizam a otimização e a explicabilidade.

1.2.1 Objetivos específicos

Os objetivos específicos são:

- Introduzir uma metodologia inovadora para a modelagem de secas que integra técnicas de aprendizado de máquina de ponta.
- Aperfeiçoar os processos de tomada de decisão, combinando modelos de aprendizado de máquina altamente otimizados com resultados transparentes e interpretáveis.
- Implementar um modelo de expansão polinomial evolucionária (EPE), aprimorado com técnicas integradas de seleção de recursos, para melhorar o desempenho do modelo.
- Implementar um modelo de expansão polinomial evolucionária (EPE), aprimorado com técnicas integradas de seleção de recursos, para melhorar o desempenho e a interpretabilidade do modelo.
- Avaliar a eficácia desses algoritmos otimizados na previsão precisa das condições de seca, fornecendo, assim, uma ferramenta confiável para o gerenciamento e planejamento de recursos.

2 REVISÃO BIBLIOGRÁFICA

Foi realizada uma revisão bibliográfica para obter uma visão abrangente da literatura de pesquisa sobre o tema. A equação lógica envolvendo os seguintes termos foi utilizada na busca nos títulos, resumos e palavras-chave dos artigos: (seca) E (previsão OU previsão OU modelagem) E (inteligência artificial OU computação suave OU aprendizado de máquina) E (índice padronizado de precipitação OU SPI). De acordo com os protocolos de triagem estabelecidos, foi conduzido um processo de revisão para avaliar os artigos. Artigos de pesquisa que não focavam em investigação baseada em dados foram excluídos. Um total de 122 artigos foi encontrado e, após um processo de triagem, 72 referências relevantes foram selecionadas, relacionadas ao tema deste estudo. A Tabela 1 apresenta o número de artigos encontrados no Scopus relacionados à previsão de seca e ao uso de aprendizado de máquina. Um aumento pode ser observado a partir de 2020, e, comparando os anos de 2023 e 2022, o número de artigos publicados dobrou. Isso mostra que o interesse em resolver o problema com maior precisão está aumentando.

Tabela 1 – Publicações por ano, 2011-2024.

Ano	**Nº de publicações**
2024	9
2023	36
2022	18
2021	16
2020	14
2019	5
2018	9
2017	3
2016	5
2015	1
2014	3
2013	2
2012	0
2011	1

A Figura 1 mostra os grafos de coocorrência de palavras-chave gerados usando o software Vosviewer. As palavras-chave foram extraídas dos corpos dos artigos e normalizadas para letras minúsculas, e as palavras de parada foram removidas. Um mínimo de 8 ocorrências nas publicações foi selecionado para clareza na representação. Sinônimos foram agrupados para evitar duplicações. As palavras-chave mais frequentes e relevantes foram selecionadas para a criação do gráfico. Um mapa de rede foi criado com base na coocorrência de palavras-chave. Os parâmetros do gráfico foram ajustados para otimizar a visualização e interpretação dos dados. Os clusters de palavras-chave no gráfico foram identificados e interpretados em relação a temas e subtemas de pesquisa. As relações entre

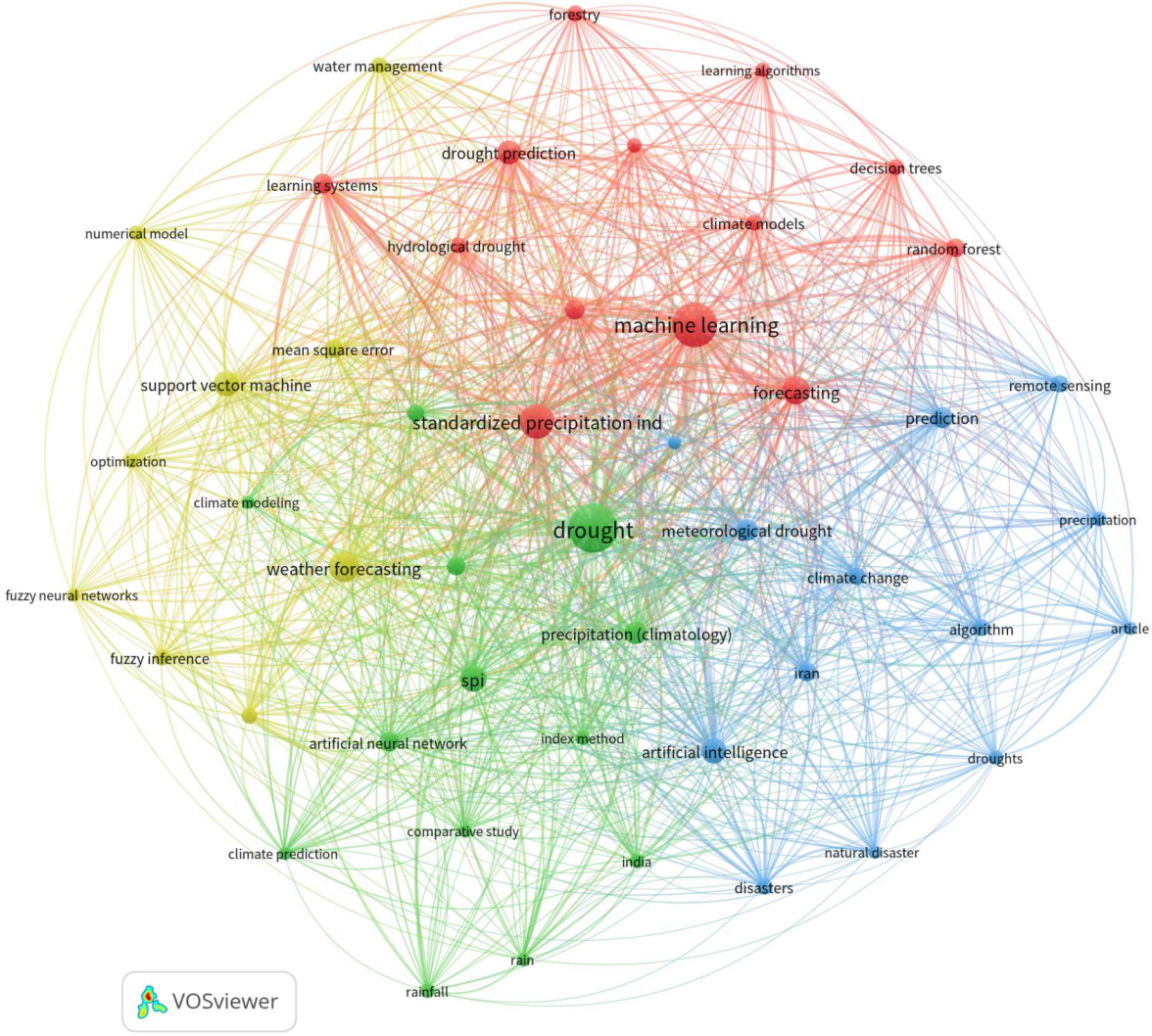


Figura 1 - Grafo de co-ocorrência de palavras-chave.

palavras-chave foram analisadas para identificar conexões entre diferentes temas e subtemas. O gráfico foi analisado para identificar tendências de pesquisa, como o surgimento de novos temas ou a evolução de temas existentes.

A Tabela 2 mostra as referências que utilizam aprendizado de máquina para prever secas e os métodos utilizados. Nota-se que a maioria dos artigos optou por usar os métodos Random Forest, Modelos Híbridos e Máquinas de Vetor de Suporte.

Tabela 2 – Métodos de aprendizado de máquina com referências

Modelo	Referência
Floresta aleatória	(12, 13, 14, 15, 16, 17, 18, 19, 20, 21, 22, 23, 24, 25, 26, 27, 28, 29, 30, 31, 32, 33, 34, 35, 36, 37, 38, 39)
Árvores de decisão	(12, 15, 22, 23, 27, 31, 40, 41)
Support Vector Machines	(14, 20, 29, 40, 42, 43, 44, 45, 46, 47, 48, 49, 50, 51, 52, 53, 54, 55, 56, 57, 58, 59, 60, 61, 62, 63, 64, 65, 66, 67, 68, 69, 70)
Fuzzy	(44, 71)
Gradient Boosting	(17, 31, 41)
Ensemble	(20, 25, 31, 36, 40, 42, 46, 53, 59, 72, 73, 10)
Modelos Híbridos	(44, 45, 47, 48, 51, 52, 62, 65, 73, 10, 74, 75, 76, 77, 78, 79, 80, 81, 82, 83)
Modelos Explicáveis	(41, 49, 84, 85)

Embora métodos explicáveis de modelagem, como a Regressão Polinomial Evolutiva (EPE), ainda não sejam amplamente abordados na literatura em comparação com técnicas de aprendizado de máquina mais populares, sua importância é crescente, especialmente em aplicações críticas como a previsão de secas e a gestão de recursos hídricos. A principal razão para adotar métodos explicáveis reside na necessidade de transparência e interpretabilidade dos modelos, permitindo que os tomadores de decisão compreendam as relações subjacentes entre as variáveis. Isso é crucial em contextos onde as decisões baseadas em modelos precisam ser justificadas, auditadas e replicáveis. Além disso, a capacidade de identificar explicitamente os fatores que influenciam os resultados do modelo pode levar a insights mais profundos e melhorias nas políticas de gestão, um aspecto frequentemente negligenciado em abordagens de "caixa-preta"como redes neurais profundas (86)

2.1 ORGANIZAÇÃO DA DISSERTAÇÃO

Esta dissertação está estruturada em cinco capítulos principais, além desta introdução. No capítulo 3, são abordados os conceitos teóricos fundamentais para o entendimento do estudo, com foco na revisão bibliográfica sobre a modelagem de secas, índices de monitoramento e as técnicas de aprendizado de máquina aplicadas na previsão de secas. O capítulo 4 descreve detalhadamente a metodologia empregada neste estudo, incluindo a área de estudo, a coleta e análise dos dados, bem como o desenvolvimento e implementação do modelo de Expansão Polinomial Evolutiva (EPE). No capítulo 5, são apresentados e discutidos os resultados obtidos a partir das simulações realizadas, destacando as expressões polinomiais geradas e a eficácia do modelo proposto. O capítulo 6 traz as conclusões

finais do trabalho, além de sugestões para trabalhos futuros, indicando caminhos para o aprimoramento das técnicas desenvolvidas e possíveis aplicações em outros contextos. Por fim, são apresentados os apêndices, que contêm pseudocódigos e outros materiais complementares que auxiliam na reprodução e compreensão dos resultados aqui discutidos.

3 FUNDAMENTAÇÃO TEÓRICA

3.1 TIPOS DE SECA

As secas podem ser classificadas em três categorias principais: meteorológica, agrícola e hidrológica, cada uma com características específicas e impactos diferenciados.

- Seca Meteorológica: Este tipo de seca é definido por um período prolongado de precipitação abaixo do normal. É geralmente o primeiro sinal de que uma seca pode estar se desenvolvendo. A seca meteorológica não leva em consideração a demanda de água, mas apenas a quantidade de precipitação recebida. Segundo (87), a seca meteorológica pode levar a condições de seca agrícola e hidrológica, dependendo da severidade e duração da deficiência de precipitação.

- Seca Agrícola: Ocorre quando a umidade do solo se torna insuficiente para sustentar as culturas agrícolas. A seca agrícola está diretamente relacionada à falta de precipitação e à evapotranspiração. Este tipo de seca afeta diretamente a produção de alimentos, podendo causar falhas nas colheitas e aumento dos preços dos alimentos, o que agrava a insegurança alimentar em regiões vulneráveis. (87) destacam que a seca agrícola pode ter consequências devastadoras para a segurança alimentar e a economia rural.

- Seca Hidrológica: Refere-se à redução dos níveis de água em corpos d'água como rios, reservatórios e aquíferos. Este tipo de seca é geralmente observado após períodos prolongados de seca meteorológica e agrícola e pode levar a sérios problemas de abastecimento de água. A seca hidrológica impacta a disponibilidade de água para consumo humano, irrigação agrícola e geração de energia hidrelétrica. Conforme discutido por (87), a seca hidrológica pode resultar em conflitos pelo uso da água e altos custos para a obtenção de água de fontes alternativas.

A figura 2, (1) ilustra como os tipos de seca estão relacionados com os fenômenos hidrológicos que por sua vez são determinados pela combinação de precipitação e umidade. Essas definições nos mostram que a precipitação não é o único fator envolvido. A intensidade das secas também é influenciada por altas temperaturas, tipo de vegetação, tipo de solo e topografia. Além disso, as atividades humanas desempenham um papel crucial, especialmente na forma como utilizamos a água e alteramos o uso do solo, como através do desmatamento e da expansão de áreas agrícolas e urbanas.

O levantamento feito por Bakshi e Shah (2) indica que pelo menos 30superfície terrestre (excluindo a Antártida) já foi afetada por secas severas e prolongadas. Entre as regiões mais impactadas estão a Califórnia, o Mediterrâneo, o Leste e o Sul da África, e

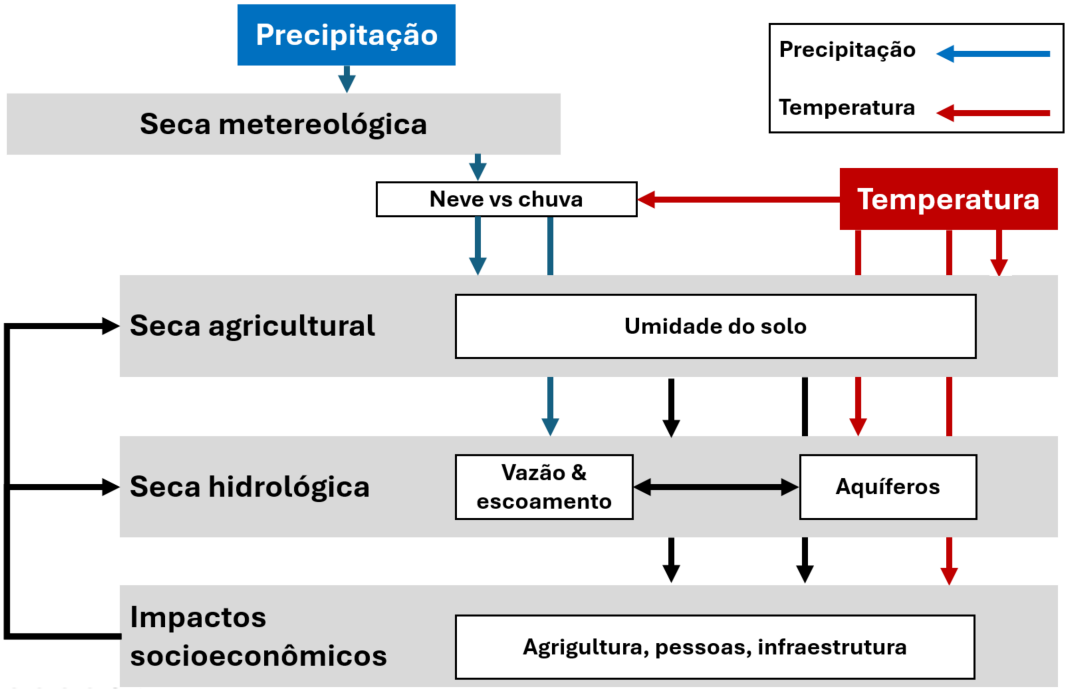


Figura 2 - Tipos de seca e sua relação com precipitação e temperatura. Adaptado de (1)

a Austrália. Segundo um estudo da NASA, cerca de 45 da superfície terrestre pode ser atingida por eventos extremos de seca até 2059.

3.2 IMPACTOS SOCIOECONÔMICOS DAS SECAS

Os impactos das secas mostrados na figura 3 são amplos e variam desde prejuízos econômicos até crises humanitárias. A seguir, são destacados alguns dos principais impactos socioeconômicos das secas segundo (3):

Agricultura e Segurança Alimentar: A seca agrícola resulta em perdas significativas de colheitas, o que reduz a produção de alimentos e aumenta a insegurança alimentar. Em regiões onde a agricultura de sequeiro é predominante, a falta de água pode levar a colheitas falhadas, forçando agricultores a abandonar suas terras e buscar meios de subsistência alternativos. Este fenômeno é especialmente crítico em países em desenvolvimento, onde a agricultura representa uma grande parte da economia e da força de trabalho.

Economia: A diminuição da produtividade agrícola durante períodos de seca pode resultar em aumentos nos preços dos alimentos, afetando a economia local e global. Além disso, as secas podem levar ao desemprego em áreas rurais, onde a agricultura é a principal fonte de emprego. As perdas econômicas associadas às secas incluem não apenas a redução da produção agrícola, mas também os custos relacionados ao racionamento de água e à importação de alimentos.

Abastecimento de Água: A seca hidrológica pode comprometer seriamente o abastecimento de água potável, afetando tanto áreas urbanas quanto rurais. A escassez de

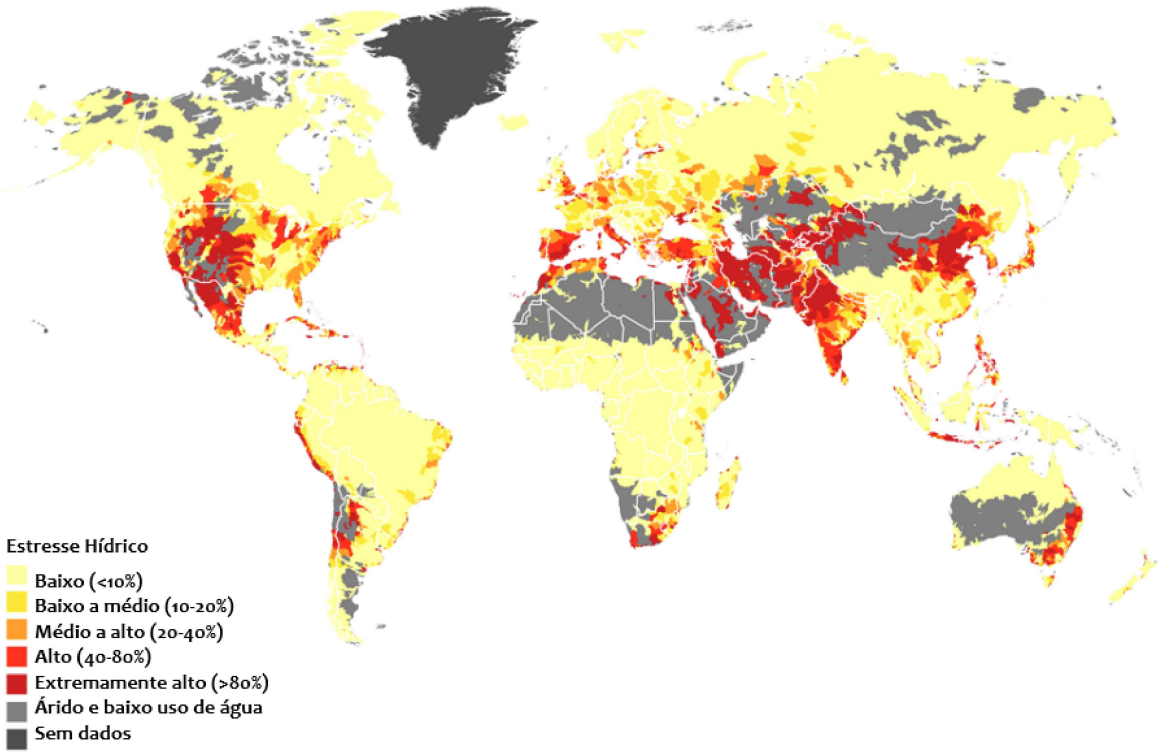


Figura 3 - Escala de estresse hídrico pelo mundo, adaptado de (2)

.

água pode levar a racionamento, conflitos pelo uso da água e altos custos para obter água de fontes alternativas. A falta de água potável também pode resultar em problemas de saúde pública, como o aumento de doenças transmitidas pela água.

Ecossistemas e Biodiversidade: A seca afeta os ecossistemas, reduzindo a disponibilidade de água para plantas e animais, o que pode levar à perda de biodiversidade. Além disso, os habitats aquáticos são particularmente vulneráveis, com quedas nos níveis de água que afetam a flora e fauna aquáticas. A redução na disponibilidade de água pode alterar os habitats naturais, colocando em risco espécies que dependem de condições hídricas específicas para sobreviver.

Conflitos e Migrações: A escassez de recursos hídricos durante períodos de seca pode levar a conflitos entre comunidades e países pelo acesso à água. Além disso, a falta de água e a perda de meios de subsistência podem forçar populações a migrar para áreas menos afetadas, criando pressões adicionais sobre os recursos dessas regiões e potencialmente exacerbando tensões sociais e políticas.

As secas são, portanto, fenômenos de grande relevância que exigem monitoramento constante e a implementação de estratégias de mitigação e adaptação para reduzir seus impactos devastadores. A utilização de índices como o *Standardized Precipitation Index (SPI)* e o *Standardised Precipitation Evapotranspiration Index (SPEI)* tem sido essencial para monitorar e analisar esses eventos extremos, permitindo uma melhor gestão dos

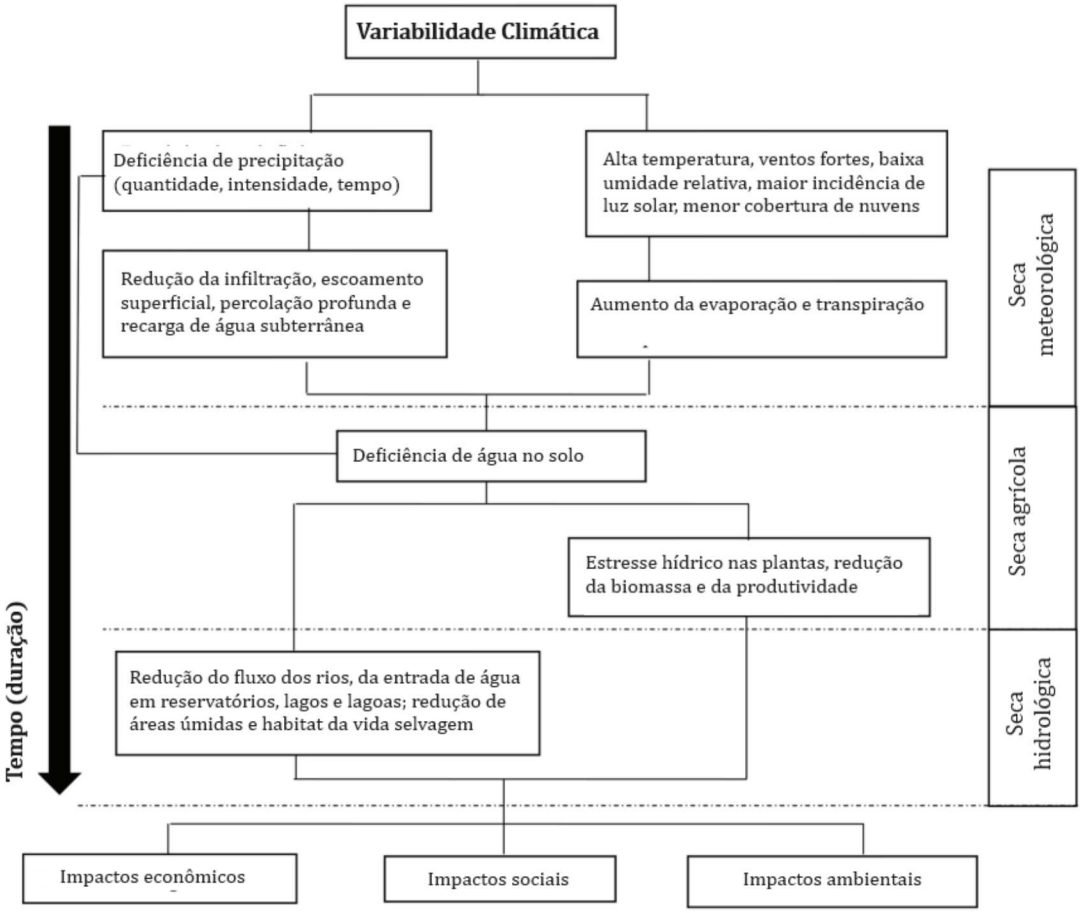


Figura 4 - Tipos de seca. Adaptado de (3)

.

recursos hídricos e a adoção de medidas preventivas.

3.3 MÉTRICAS PARA MONITORAMENTO DE SECAS

Para estimar a gravidade da seca, os indicadores relativos são medidos utilizando um conjunto de informações sobre meteorologia, agricultura, hidrologia ou socioeconomia. Fatores climáticos, como temperatura e precipitação, são frequentemente aplicados para determinar o índice de seca. Esses índices avaliam a severidade da seca em desenvolvimento e, se apoiados pela estrutura institucional adequada, podem ser utilizados para acionar planos de contingência contra a seca (3).

Alguns indicadores de seca baseados na precipitação foram propostos, como o índice padronizado de precipitação (SPI) e o índice de seca efetivo (EDI). No entanto, embora a observação, apesar de funcional para determinar sinais precoces de secas, identifique apenas casos já ocorrendo. O desafio é estimar os futuros dias, semanas, meses de seca e sua intensidade, ou seja, aumentar o potencial de alerta antecipado dos sistemas de monitoramento de seca por meio da previsão de seca. Embora esses índices sejam

inicialmente meteorológicos, sua medida seria instrutiva na administração da seca agrícola e no desempenho dos recursos hídricos.

3.3.1 Índice padronizado de precipitação

O Índice de Precipitação Padronizado (SPI), proposto por (5), é uma medida amplamente utilizada para quantificar a deficiência de precipitação em diversas escalas temporais. O SPI permite a caracterização das secas com base em dados históricos de precipitação, oferecendo uma forma objetiva de identificar e monitorar eventos de seca.

3.3.2 Definição e Cálculo do SPI

O SPI é definido como a diferença entre a precipitação acumulada durante um período específico e a média histórica desse período, normalizada pela variabilidade da precipitação (desvio padrão). A fórmula matemática do SPI é dada pela equação 3.1:

$$\text{SPI} = \frac{X_i - \overline{X}}{\sigma} \tag{3.1}$$

onde:

- X_i é a precipitação acumulada no i-ésimo período.
- $\overline{X}$ é a média histórica da precipitação para o mesmo período.
- σ é o desvio padrão da precipitação histórica.

3.3.3 Metodologia de Cálculo do SPI

1. **Coleta de Dados**: Um conjunto de dados mensais de precipitação é preparado para um período de m meses, idealmente abrangendo pelo menos 30 anos de registros contínuos.
2. **Escalas Temporais**: Seleciona-se um conjunto de períodos de média j meses, onde j pode ser 3, 6, 12, 24 ou 48 meses. Essas escalas representam diferentes tipos de fontes de água utilizáveis, como água do solo, águas subterrâneas, escoamento superficial e armazenamento em reservatórios.
3. **Ajuste à Função Gama**: Cada conjunto de dados é ajustado à distribuição Gamma para definir a relação de probabilidade com a precipitação. A função Gama é utilizada devido à sua flexibilidade em modelar a distribuição assimétrica típica da precipitação.
4. **Cálculo da Probabilidade**: A probabilidade de qualquer ponto de dado observado de precipitação é calculada e utilizada, juntamente com uma estimativa da inversa

normal, para calcular a desvio da precipitação para uma densidade de probabilidade normalmente distribuída com média zero e desvio padrão unitário.

5. **Cálculo do SPI**: O valor do SPI é então calculado para o ponto de dado específico de precipitação.

3.3.4 Categorias de Seca

O trabalho de (5), também propôs a definição de categorias de seca baseadas nos valores do SPI, permitindo uma classificação quantitativa da intensidade da seca como pode ser visto na tabela 3:

Tabela 3 – Categorias de Seca baseadas no valor do SPI

Valor do SPI	**Categoria de Seca**
0 a -0.99	Seca Leve
-1.00 a -1.49	Seca Moderada
-1.50 a -1.99	Seca Severa
$\leq$ -2.00	Seca Extrema

O SPI apresenta diversas vantagens em relação a outros índices de seca:

- **Normalização**: O SPI é normalizado, permitindo comparações entre diferentes regiões climáticas.
- **Flexibilidade**: Pode ser calculado para diferentes escalas temporais, tornando-o aplicável a diversas fontes de água utilizáveis.
- **Probabilidade**: Está diretamente relacionado à probabilidade de ocorrência de eventos de seca, facilitando a interpretação e a comunicação dos resultados.

A introdução do SPI por (5) forneceu uma ferramenta robusta e versátil para a monitorização e análise de secas, reconhecendo a importância das escalas temporais e proporcionando um método padronizado para a avaliação da deficiência de precipitação. A aplicação do SPI em diversas regiões tem demonstrado sua eficácia na gestão de recursos hídricos e na mitigação dos impactos da seca.

3.3.5 Índice de Seca Efetivo (EDI)

O EDI considera a precipitação diária e o déficit de água no solo, incorporando a evapotranspiração potencial. Este índice é calculado com base no déficit acumulado de precipitação efetiva, que é a precipitação diária menos a evapotranspiração potencial. O EDI reflete a resposta imediata do sistema hídrico à deficiência de precipitação. As principais aplicações do EDI incluem:

- Avaliação da seca agrícola e hidrológica.
- Monitoramento de curto prazo, com rápida resposta às mudanças nas condições de seca.
- Gestão de irrigação e suporte à tomada de decisão em tempo real.

3.3.6 Principais Diferenças entre SPI e EDI

Ao analisar os pontos relevantes do SPI segundo (5) e EDI segundo (8) é possível fazer a seguinte comparação:

- Escala Temporal: O SPI pode ser calculado para várias escalas temporais (mensal, trimestral, anual, etc.), enquanto o EDI foca nas mudanças diárias, sendo mais sensível às variações de curto prazo.
- Dados Utilizados: O SPI utiliza precipitação acumulada em períodos mais longos, ao passo que o EDI baseia-se na precipitação diária e na evapotranspiração potencial.
- Sensibilidade: O EDI responde mais rapidamente às mudanças nas condições de seca devido à sua base diária, enquanto o SPI oferece uma visão mais estabilizada ao longo de períodos maiores.
- Aplicações Específicas: O SPI é mais adequado para análises de longo prazo e comparações regionais de secas, ao passo que o EDI é útil para monitoramento em tempo real e decisões imediatas relacionadas à agricultura e gestão de água.

Em resumo, ambos os índices são valiosos para a gestão e monitoramento de secas, mas são aplicáveis em diferentes contextos e escalas temporais, complementando-se mutuamente para fornecer uma visão abrangente das condições de seca.

3.4 INTELIGÊNCIA ARTIFICIAL EXPLICÁVEL

No contexto do estudo abordado na dissertação, a Inteligência Artificial Explicável (XAI) desempenha um papel crucial na modelagem e previsão de secas, segundo (84), principalmente ao balancear precisão preditiva e transparência interpretativa. A XAI permite que modelos complexos, como o Expansão Polinomial Evolutiva (EPE) utilizado neste estudo, sejam compreendidos não apenas em termos de acurácia, mas também em termos dos fatores subjacentes que influenciam as previsões. Ao integrar técnicas de seleção de características, como a Lasso, o modelo EPE não só melhora a eficiência computacional, mas também promove uma compreensão clara das relações entre variáveis climáticas e o índice de precipitação padronizado (SPI), fundamental para a previsão de secas em áreas vulneráveis.

A relevância da XAI em contextos hidrológicos e climatológicos, como a previsão de secas, reside na sua capacidade de fornecer modelos que são tanto precisos quanto interpretáveis. No caso do modelo EPE, a capacidade de derivar expressões polinomiais que elucidam as interações entre variáveis climáticas oferece aos gestores de recursos hídricos uma ferramenta valiosa para decisões informadas. Essas expressões interpretáveis, derivadas através de aprendizado evolutivo, permitem uma melhor avaliação e compreensão dos fatores que contribuem para eventos de seca, o que é crucial para o desenvolvimento de estratégias de mitigação eficazes.

Além disso, o uso de XAI no contexto de previsão de secas representa um avanço significativo na capacidade de modelos preditivos de serem auditados e validados por especialistas. Ao possibilitar uma análise detalhada dos componentes do modelo e das relações entre as variáveis, a XAI aumenta a confiança nas previsões, facilitando a adoção de tais modelos em políticas públicas e estratégias de planejamento. Estudos como os de (88) e (84) destacam a importância de combinar modelos de aprendizado de máquina com abordagens explicáveis para melhorar tanto a transparência quanto a aplicabilidade desses modelos em cenários complexos e críticos

3.5 MODELO DE EXPANSÃO POLINOMIAL EVOLUTIVA (EPE)

Este trabalho, utiliza o modelo de Expansão Polinomial Evolutiva (EPE) com seleção de características Lasso, uma abordagem para a previsão do índice SPI que prioriza a interpretabilidade e aborda a complexidade dos dados. Ao contrário dos modelos tradicionais de *caixa preta*, o EPE aproveita um mecanismo de aprendizado evolutivo que se baseia em expressões polinomiais interpretáveis. No entanto, os dados hidrometeorológicos podem ser ricos em características, algumas potencialmente irrelevantes ou redundantes. É aqui que a seleção de características Lasso desempenha um papel importante (86).

Esta abordagem integra o Lasso, uma técnica conhecida por reduzir coeficientes e promover a esparsidade, no processo de aprendizado do modelo EPE. O Lasso analisa efetivamente as características dos dados e identifica as mais relevantes que contribuem significativamente para a previsão de fluxos de água. A incorporação do Lasso permite que o modelo EPE se concentre na construção de sua estrutura interna utilizando apenas as características mais impactantes. Isso não só melhora a eficiência do modelo, mas também aumenta a interpretabilidade. O modelo resultante prevê o SPI e destaca os principais fatores expressos através de funções polinomiais interpretáveis que impulsionam essas previsões. Esse nível de explicabilidade capacita pesquisadores e gestores de recursos hídricos a obter insights valiosos sobre os fatores críticos do comportamento do fluxo de água. Usando o conhecimento descoberto, os especialistas podem tomar decisões mais informadas para a gestão de recursos hídricos e estratégias de mitigação de secas.

3.6 MODELO DE REGRESSÃO POLINOMIAL EVOLUTIVA (EPR)

A Regressão Polinomial Evolutiva (EPR) é uma técnica de modelagem que combina a robustez dos algoritmos genéticos com a simplicidade e interpretabilidade dos modelos polinomiais. Desenvolvida como uma ferramenta para capturar interações não lineares e complexas em sistemas de engenharia e fenômenos naturais, o EPR permite a construção de modelos que são ao mesmo tempo precisos e facilmente interpretáveis. A abordagem se destaca por sua capacidade de selecionar automaticamente as variáveis mais relevantes e de encontrar a estrutura polinomial que melhor se ajusta aos dados observacionais, minimizando erros de previsão. O processo evolutivo utilizado no EPR, que combina algoritmos genéticos com métodos de mínimos quadrados, garante que os modelos gerados sejam otimizados tanto em termos de precisão quanto de simplicidade, evitando o overfitting e promovendo a generalização dos resultados (11).

Comparando as duas técnicas similares, é possivel observar que o EPE Oferece maior interatividade ao usuário, permitindo que ele faça suposições sobre a estrutura do modelo e sobre quais elementos devem ser considerados na função alvo. Isso oferece mais controle sobre o processo de modelagem. Já para o EPR, embora o usuário ainda tenha controle sobre certos parâmetros, como o número de termos e o tipo de função, o foco está mais na automatização da seleção do modelo através do processo evolutivo.

3.6.1 Engenharia de Características via Expansão Polinomial

A engenharia de recursos desempenha um papel crucial na extração de características informativas dos dados brutos e na melhoria do desempenho dos modelos de aprendizado de máquina. A expansão polinomial é uma técnica valiosa de engenharia de recursos para previsão de secas usando o modelo EPE.

Os dados hidrometeorológicos frequentemente exibem relações complexas entre várias variáveis climáticas que influenciam a severidade da seca. Abordagens tradicionais podem ter dificuldade em capturar essas relações não-lineares de forma eficaz (9). A expansão polinomial aborda essa limitação gerando novos recursos com base nos existentes através de suas interações binárias e de ordem superior.

Sejam $x_1, x_2, ..., x_d$ as características originais no conjunto de dados. A expansão polinomial visa criar um novo conjunto de características, F_p, através da expressão 3.2:

$$F_p = \{f_k(x_i, x_j) \mid i, j \in \{1, 2, ..., D\}, i < j, k \in \{1, 2, ..., K\}\} \cup \{x_i^2, ..., x_d^p\} \tag{3.2}$$

onde i e j são índices iterando pelas características originais, as funções $f_k(x_i, x_j)$ representam a criação de novos recursos através de interações binárias entre as características originais, e K é o número de funções de expansão. O segundo termo, $x_i^2, ..., x_d^p$, introduz as características originais.

Neste estudo, K foi definido como 3, levando a:

$$F_p = \{f_1(x_i, x_j), f_2(x_i, x_j), f_3(x_i, x_j)\} \tag{3.3}$$

e as seguintes funções de expansão:

$$f_1(x_i, x_j) = x_i \cdot x_j \tag{3.4}$$

$$f_2(x_i, x_j) = \frac{x_i}{x_j} \tag{3.5}$$

$$f_3(x_i, x_j) = \frac{x_j}{x_i} \tag{3.6}$$

estabelecem diferentes interações entre x_i e x_j. Como todas as observações do SPI são positivas, as funções são definidas para quaisquer valores de x_i e x_j, permitindo expansões em todos os termos. Nos casos em que as observações podem assumir valores próximos de zero, as expansões devem ser calculadas com cuidado para evitar indeterminações.

Considerando n variáveis, cada função $f_i(x_i, x_j)$ expande em $n(n-1)/2$ termos, e usando a formulação na Equação 3.1, o número de termos no espaço de funções expandidas é dado por $n + Kn(n-1)/2$, onde n é a dimensão do espaço de variáveis originais e K é o número de funções de expansão.

3.7 INTEGRAÇÃO COM O MODELO EPE

O modelo EPE aproveita o conjunto de características expandidas gerado através da expansão polinomial. Durante o processo de aprendizado evolutivo, o modelo seleciona e combina dinamicamente esses novos recursos juntamente com os originais para construir sua estrutura interna. Isso permite que o modelo EPE identifique as interações mais relevantes e as relações não-lineares entre variáveis climáticas que influenciam os valores do SPI, melhorando, em última análise, a precisão da previsão de secas.

O Lasso minimiza a soma residual dos quadrados sujeita à soma do valor absoluto dos coeficientes ser menor que uma constante (46). Devido à natureza dessa restrição, ele tende a produzir alguns coeficientes que são exatamente 0 e, portanto, gera modelos interpretáveis.

A Evolução Diferencial (DE) é um algoritmo de otimização estocástica baseado em uma população de soluções, que opera através de etapas computacionais semelhantes às empregadas pela maioria dos Algoritmos Evolutivos. DE é simples e fácil de implementar, e possui algumas características como desempenho superior em relação à precisão, velocidade de convergência e robustez, e poucos parâmetros de controle.

A função objetivo, a ser minimizada pela evolução diferencial, é escrita como segue:

$$f(x) = \text{RMSE}(x) \times \left(1 + \frac{\beta}{N_{\text{feat}}} \sum_{i=1}^{N_{\text{feat}}} x_i^{\text{MB}}\right) \tag{3.7}$$

O parâmetro β controla indiretamente a complexidade das expressões geradas para o modelo linear. Dessa equação, observa-se que quando $\beta = 0$, a expressão resultante da solução candidata não é penalizada por seu comprimento. À medida que β aumenta, expressões com mais termos são mais penalizadas do que expressões com poucos termos. Uma escolha adequada de β é crucial para equilibrar a precisão da expressão e a interpretabilidade do modelo de previsão associado.

3.8 VALIDAÇÃO CRUZADA DE SÉRIES TEMPORAIS

O uso de modelos de aprendizado de máquina em dados de séries temporais requer cuidados especiais devido às suas peculiaridades. Deve-se escolher o conjunto de teste dentro de um período de tempo após o conjunto de treinamento. Caso contrário, algumas informações podem vazar do conjunto de treinamento para o conjunto de teste, comprometendo o processo de aprendizado do modelo de aprendizado de máquina. Uma solução mais robusta é operar de forma semelhante à validação cruzada *k-fold*, mas de maneira ordenada no tempo (89). A Figura 5 ilustra o procedimento de validação cruzada para séries temporais. Podemos obter uma estimativa imparcial do desempenho do modelo treinando e ajustando o modelo no conjunto de treinamento para cada fold e calculando a média dos erros nos conjuntos de teste.

Fold 1	Teste	Treino	Treino	Treino
Fold 2	Treino	Teste	Treino	Treino
Fold 3	Treino	Treino	Teste	Treino
Fold 4	Treino	Treino	Treino	Teste

Figura 5 - Esquema de validação cruzada k-folds com 4 folds.

3.9 MÉTRICAS DE DESEMPENHO

O modelo é avaliado usando métricas apropriadas ao tipo de problema, como a eficiência Kling-Gupta (KGE), erro absoluto médio (MAE), erro absoluto percentual médio (MAPE), erro quadrático médio (MSE), coeficiente de eficiência de Nash-Sutcliffe (NSE), erro quadrático médio da raiz (RMSE), erro quadrático médio relativo (RRMSE), coeficiente de determinação (R^2) e índices de Willmott (WI) são usado (9).

$$R\check{s} = 1 - \frac{\sum_i (y_i - \overline{y})\check{s}}{\sum_i (f_i - \overline{y})\check{s}} \tag{3.8}$$

$$WI = \frac{1}{N}\sum_{i=1}^{N}|y_i - \hat{y}_i| \tag{3.9}$$

$$RRMSE = \sqrt{\frac{\sum_{i=1}^{N}(y_i - \hat{y}_i)\check{\text{s}}}{\sum_{i=1}^{N} y_i\check{\text{s}}}} \tag{3.10}$$

$$RMSE = \sqrt{\frac{1}{N}\sum_{i=0}^{N-1}(y_i - \hat{y}_i)\check{\text{s}}} \tag{3.11}$$

$$MAE = \frac{\sum_{i=1}^{N}|y_i - \hat{y}_i|}{N} \tag{3.12}$$

$$MAPE = \frac{1}{N}\left(\sum_{i=1}^{N}\frac{|y_i - \hat{y}_i|}{|y_i|}\right)100 \tag{3.13}$$

$$NSE = 1 - \frac{\sum_{i=1}^{N}(y_i - \hat{y}_i)\check{\text{s}}}{\sum_{i=1}^{N}(y_i - \overline{y})\check{\text{s}}} \tag{3.14}$$

$$KGE = 1 - \sqrt{(r-1)\check{\text{s}} + (\alpha - 1)\check{\text{s}} + (\gamma - 1)\check{\text{s}}} \tag{3.15}$$

Os valores reais e simulados nas equações são representados por y e $\hat{y}$, respectivamente. $\overline{y}$ representa a média real e r é o coeficiente de Pearson e α indica a proporção entre os desvios padrão dos valores simulados e reais. Finalmente, γ é a proporção entre as médias dos valores simulados e observados.

4 METODOLOGIA

4.1 ÁREA DE ESTUDO

Os dados utilizados, obtidos de Ali (83) foram coletados em estações meteorológicas na região de Ancara, capital da Turquia no período de 1971 até 2015. Esta cidade possui uma extensão territorial de 24.521 km^2 e mais de 5,5 milhões de habitantes. A precipitação média anual é de 383 mm, e a temperatura média é de 11,6°C. Medidas históricas mensais de precipitação das estações meteorológicas de Beypazarı, Esenboğa, Nallıhan, Keçiören, Kızılcahamam e Polatlı foram utilizadas para construir o modelo. As localizações dessas estações na região estão ilustradas na Figura 6. A Tabela 4 mostra a associação de cada variável com a estação em que os experimentos foram conduzidos.

Beypazarı está situada a oeste de Ancara e é conhecida por suas atividades agrícolas e importância histórica. O clima de Beypazarı, caracterizado por condições semiáridas, faz dela um indicador essencial para a modelagem de secas. Os padrões de precipitação em Beypazarı podem fornecer sinais antecipados de mudanças nas condições meteorológicas que podem afetar a região mais ampla, incluindo Ancara. Como um centro agrícola, as mudanças na precipitação em Beypazarı podem ter implicações significativas para o uso e disponibilidade de água, impactando tanto a agricultura local quanto os recursos hídricos em Ancara.

Polatlı é outro centro agrícola localizado a oeste de Ancara. Seus dados de precipitação são vitais para entender a disponibilidade regional de água, especialmente considerando suas extensas atividades agrícolas. As condições climáticas em Polatlı podem influenciar as estratégias de gestão de água e a preparação para secas em Ancara.

Kızılcahamam, localizada ao norte de Ancara, apresenta um terreno montanhoso e áreas florestais que experimentam padrões de precipitação diferentes em comparação com as planícies centrais. Os efeitos orográficos da área e as variações climáticas locais contribuem com dados climáticos únicos que aumentam a precisão do modelo na previsão de condições de seca em Ancara. A precipitação em Kızılcahamam influencia o balanço hídrico geral da região.

Esenboğa, localizada a nordeste de Ancara, é uma estação meteorológica crítica devido à sua proximidade com o principal aeroporto de Ancara. Esta área fornece dados climáticos de alta qualidade em tempo real, cruciais para a modelagem precisa de secas.

Nallıhan, situada a noroeste de Ancara, possui uma mistura de terras agrícolas e reservas naturais. Os dados de precipitação de Nallıhan capturam padrões climáticos únicos que outras estações podem não registrar, proporcionando uma visão abrangente das influências climáticas regionais nas condições de seca em Ancara.

Keçiören, localizada na área urbana de Ancara, fornece dados valiosos sobre

padrões de precipitação urbana. Como um distrito densamente populado, as medições de precipitação de Keçiören são cruciais para entender o impacto da urbanização no clima local e nos recursos hídricos.

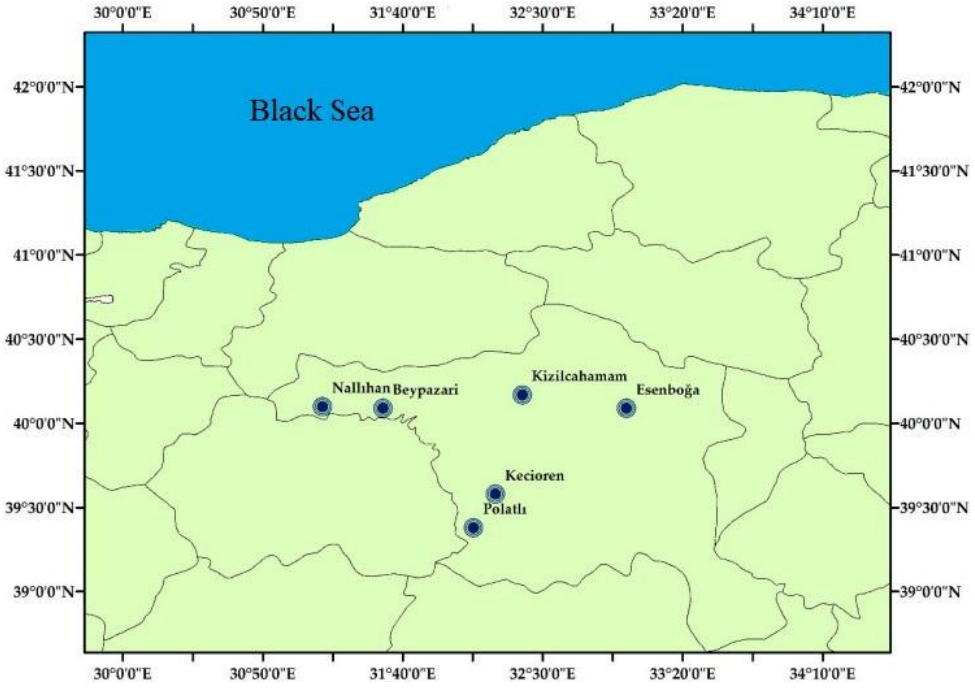
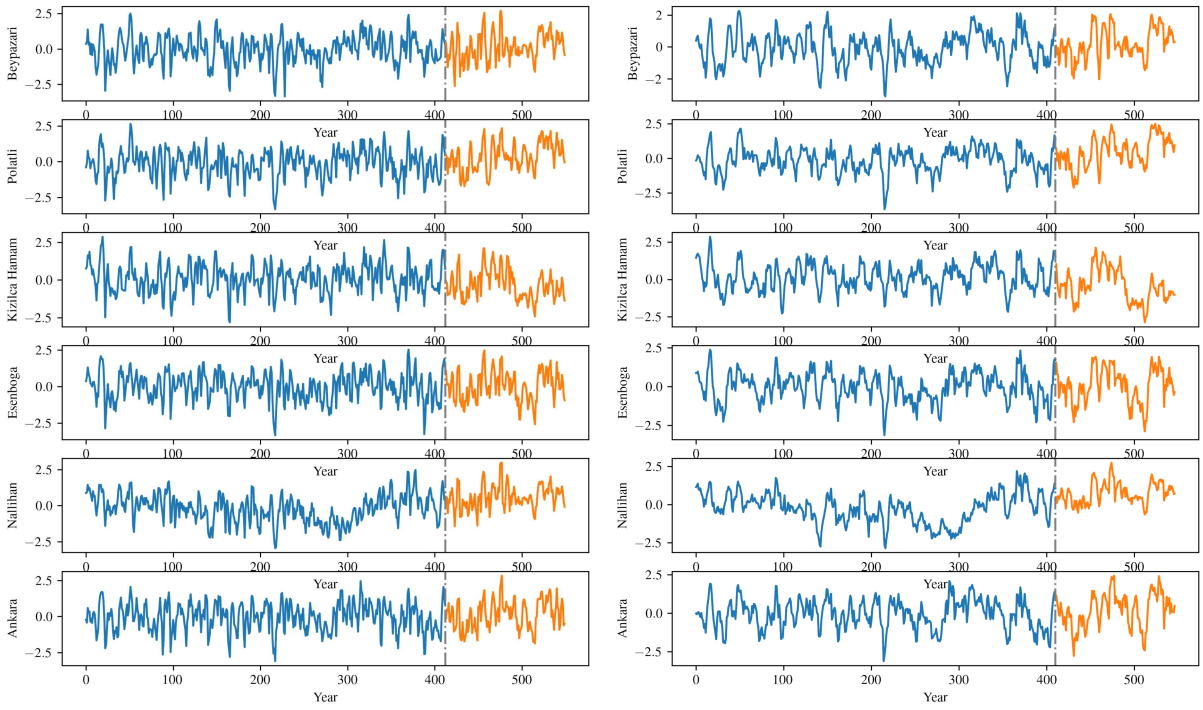


Figura 6 - Localização das seis estações meteorológicas na cidade de Ancara, capital da Turquia, utilizadas no estudo. Adaptado de (4).

Tabela 4 – Associação entre estações e variáveis.

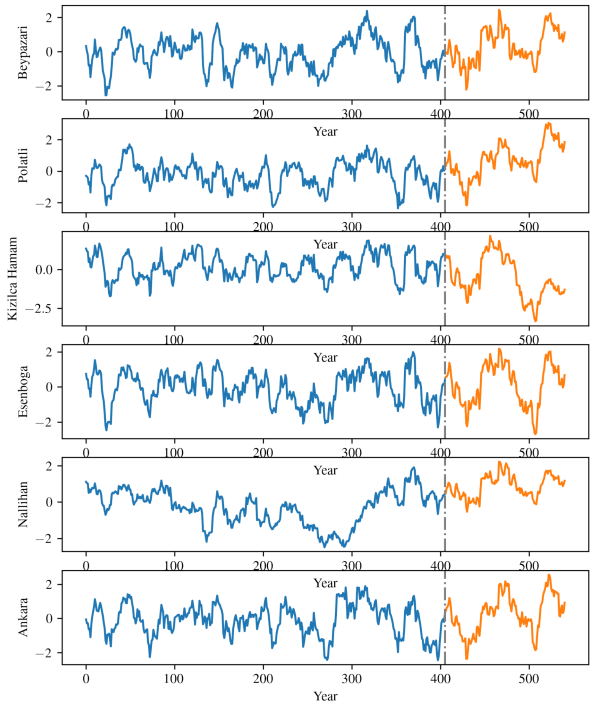
Estação	Variável
Beypazari	x_1
Polatli	x_2
Kizilca Hamam	x_3
Esenboga	x_4
Nallihan	x_5
Kecioren	x_6
Ancara	y

A análise de previsão do índice SPI será realizada utilizando dados de medições anteriores. A Figura 7 mostra as medições históricas de precipitação para as variáveis de entrada das estações (azul) e saída (laranja) para os 3, 6 e 12 meses anteriores, respectivamente.



(a)

(b)



(c)

Medições históricas de precipitação mensal para o Índice Padronizado de Precipitação (a) SPI-3, (b) SPI-6, e (c) SPI-12. O conjunto de treino é indicado em azul e do conjunto de teste em laranja.

4.2 ANÁLISE EXPLORATÓRIA DOS DADOS

Os experimentos computacionais foram realizados utilizando a linguagem Python e suas bibliotecas pandas (90), scipy (91) e o framework scikit-learn (92).

A análise exploratória de dados (EDA) do Índice de Precipitação Padronizada (SPI) desempenha um papel fundamental na compreensão dos padrões de seca e na gestão de recursos hídricos. A EDA permite identificar e visualizar características temporais e espaciais das secas, facilitando a identificação de padrões e anomalias que podem informar estratégias de mitigação e planejamento de recursos hídricos.

O estudo conduzido por (93) proporciona uma análise abrangente dos períodos de seca agrícola e hidrológica no Paquistão. A pesquisa se concentrou nas escalas de tempo de 3 e 12 meses do SPI, conhecidas por refletirem mudanças sazonais e hidrológicas na intensidade da seca, respectivamente.

Uma parte crucial da EDA é a utilização de gráficos de séries temporais, que permitem visualizar as flutuações do SPI ao longo do tempo. Esses gráficos ajudam a identificar períodos de seca severa e sua duração, além de fornecer *insights* sobre a frequência e a intensidade das secas em diferentes locais. Tabelas detalhadas complementam esses gráficos, fornecendo uma visão resumida dos períodos de seca, incluindo dados sobre início, término, duração e valores de pico de intensidade. A figura 8 mostra os dados de SPI das estações do conjunto de dados de SPI-12, destacando os períodos de seca conforme a definição de (5)

A análise de séries temporais desempenha um papel crucial nos métodos de previsão do Índice Padronizado de Precipitação (SPI), uma métrica amplamente utilizada para quantificar anomalias de precipitação e monitorar eventos de seca. Através da decomposição de séries temporais, é possível identificar componentes fundamentais como tendência, sazonalidade e ruído, cada um dos quais oferece insights valiosos sobre o comportamento dos dados ao longo do tempo. A sazonalidade, em particular, é uma característica essencial na análise de séries temporais de precipitação, dado que padrões sazonais repetitivos podem fornecer informações importantes para a seleção de features na modelagem preditiva. A identificação de ciclos sazonais permite ajustar modelos que capturam variações periódicas na precipitação, melhorando assim a precisão das previsões do SPI. Estudos como o de (94) , que explorou a aplicação do SPI em diversas escalas temporais, e o trabalho de (5), que introduziu o conceito do SPI, destacam a importância de considerar a sazonalidade nos processos de modelagem.

A compreensão detalhada da sazonalidade é vital, pois esta pode variar significativamente de uma região para outra, influenciada por fatores climáticos locais e padrões meteorológicos globais. Ao identificar corretamente esses padrões sazonais, os modelos preditivos podem ajustar seus parâmetros de maneira a capturar melhor as flutuações

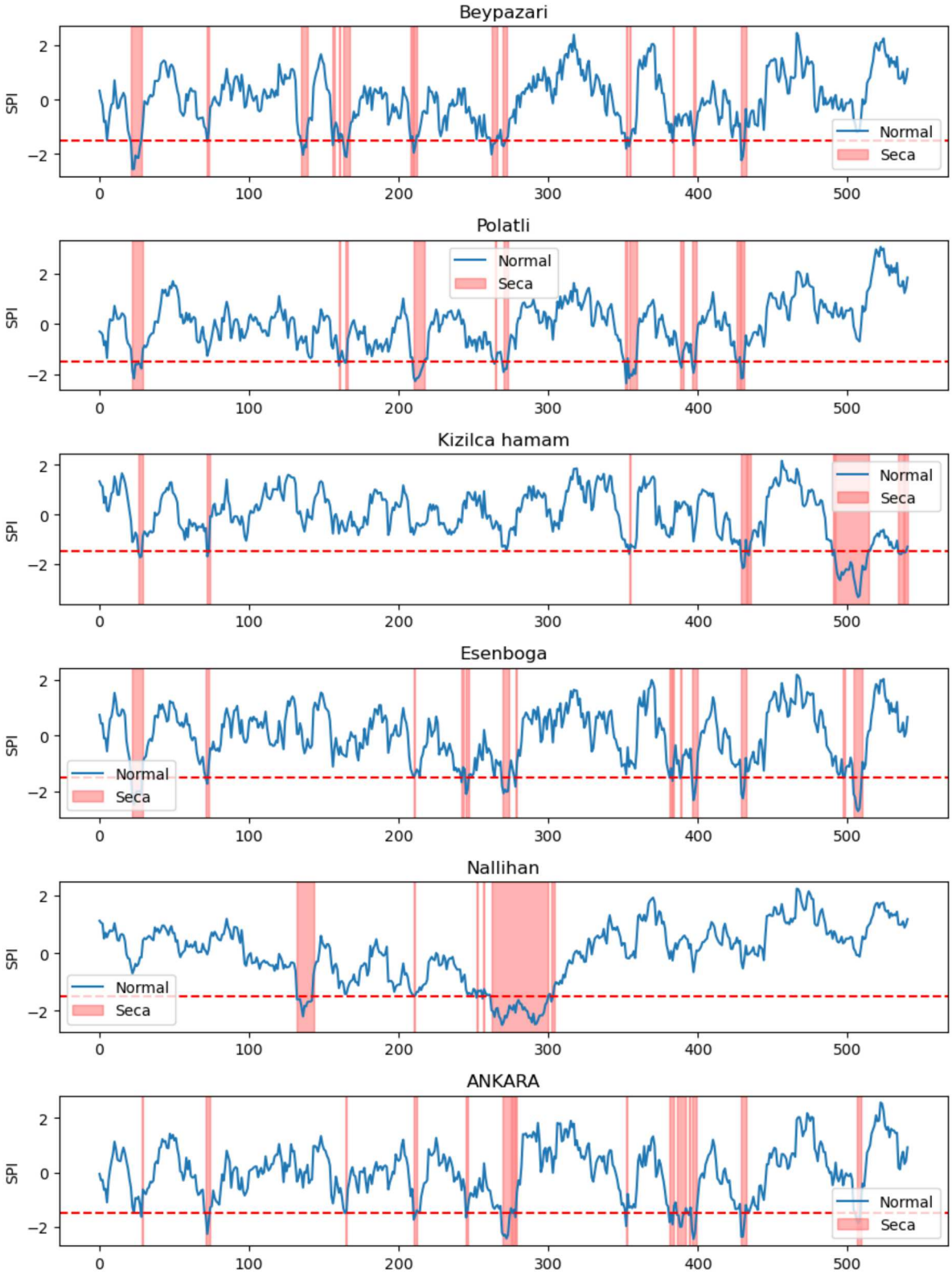


Figura 8 - Períodos de seca do conjunto de dados segundo a clasificação de (5)

esperadas na precipitação, reduzindo assim os erros de previsão. Por exemplo, regiões tropicais podem apresentar uma sazonalidade marcada por estações chuvosas e secas bem definidas, enquanto regiões temperadas podem mostrar variações sazonais ligadas a

mudanças de temperatura e precipitação ao longo do ano. Esta compreensão facilita a incorporação de variáveis que capturam essas nuances sazonais nos modelos, aprimorando a capacidade de previsão.

Além disso, a análise de séries temporais permite a detecção de tendências de longo prazo na precipitação, que podem ser atribuídas a mudanças climáticas ou a outros fenômenos de escala global. A identificação de tendências e sua separação dos componentes sazonais e aleatórios é essencial para prever mudanças futuras na disponibilidade de água e para planejar estratégias de mitigação de secas. Ferramentas como o modelo ARIMA (*AutoRegressive Integrated Moving Average*) ou técnicas mais avançadas de machine learning, como redes neurais recorrentes (RNNs) e LSTM (Long Short-Term Memory), têm sido aplicadas com sucesso na modelagem de séries temporais de precipitação, incorporando sazonalidade e tendências para melhorar a precisão preditiva (39).

A integração de análise de séries temporais com técnicas de aprendizado de máquina pode ainda melhorar a capacidade de previsão do SPI. Métodos híbridos que combinam decomposição de séries temporais com algoritmos de aprendizado supervisionado têm mostrado resultados promissores. Por exemplo, a decomposição da série temporal pode ser utilizada para extrair componentes sazonais e de tendência, que são então usados como inputs para modelos de machine learning, como suporte vector machines (SVM) ou redes neurais artificiais (ANN). Esta abordagem permite que o modelo aprenda padrões complexos nos dados que podem não ser capturados por métodos tradicionais (80).

Estudos recentes têm explorado a utilização de redes neurais convolucionais (CNNs) para a análise de séries temporais climáticas, aproveitando sua capacidade de capturar padrões espaciais e temporais nos dados. A pesquisa de (95) demonstrou que a combinação de CNNs com LSTMs pode melhorar significativamente a previsão do SPI, capturando tanto as variações sazonais quanto as tendências de longo prazo. Essa abordagem híbrida exemplifica como a análise de séries temporais, quando integrada com técnicas avançadas de aprendizado de máquina, pode proporcionar previsões mais precisas e robustas.

Em resumo, a análise de séries temporais é fundamental na previsão do Índice Padronizado de Precipitação devido à sua capacidade de decompor a série em componentes interpretáveis, como tendência, sazonalidade e ruído. A identificação e modelagem da sazonalidade permitem uma seleção de features mais informada e ajustada, o que é crucial para a precisão dos modelos preditivos. A literatura, incluindo os trabalhos de (96) e (5), enfatiza a importância dessa abordagem para a melhoria das previsões do SPI. Além disso, a combinação de métodos de séries temporais com técnicas de aprendizado de máquina, como mostrado por (14), pode levar a avanços significativos na previsão de eventos de seca, oferecendo ferramentas mais eficazes para a gestão de recursos hídricos e mitigação de desastres naturais.

A matriz de correlação mostrada na Figura 9 é outra ferramenta essencial na EDA,

pois ajuda a identificar as relações entre diferentes variáveis climáticas e o SPI. Ao analisar as correlações, é possível determinar quais fatores climáticos têm maior influência nas secas, o que é crucial para desenvolver modelos preditivos mais precisos.

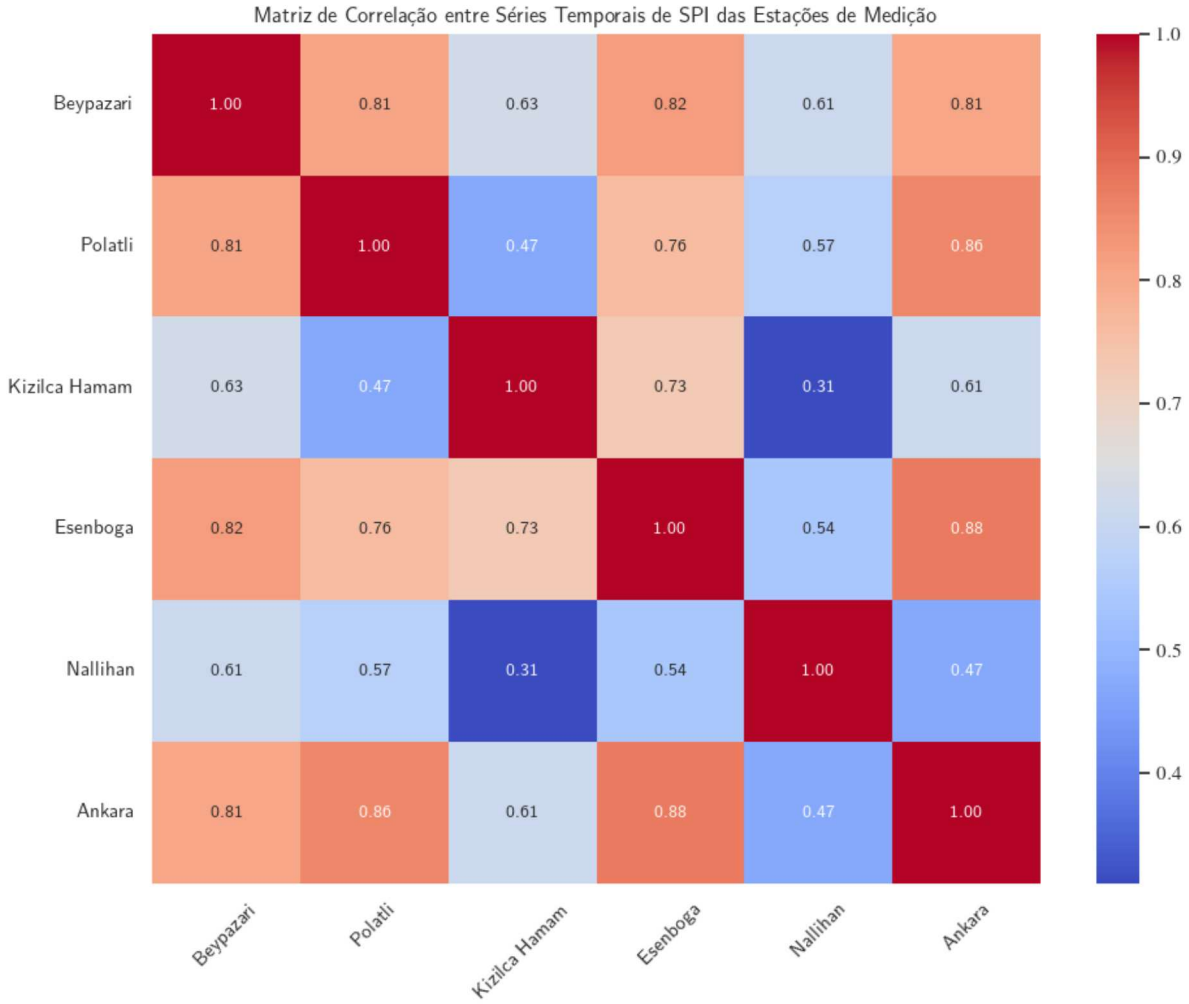


Figura 9 - Matriz de correlação.

A análise da função de autocorrelação (ACF) e da função de autocorrelação parcial (PACF) também desempenha um papel significativo na EDA. A ACF permite entender a dependência temporal dos dados, mostrando como os valores do SPI estão correlacionados com seus valores passados. Já a PACF ajuda a identificar a ordem dos modelos de séries temporais, indicando quantos lags (atrasos) são relevantes para prever o SPI. Essas análises são fundamentais para modelar corretamente as séries temporais e prever eventos de seca com maior precisão. A Figura 10 mostra os gráficos de auto correlação total e parcial para os dados das estações. É possível observar que, devido ao comportamento sazonal dos fenômenos hidrológicos, há forte correlação parcial em intervalos de 12 meses se repetindo em menor escala em um intervalo de 24 meses. Esses intervalos estão marcados pela linha tracejada do gráfico. Além disso, percebe-se que existe uma correlação com os pontos anteriores no conjunto de dado, um ponto em um instante t possui uma correlação alta

com os pontos em $t-1$ o que diminuí até o lag $t-12$. Esse comportamento do conjunto de dados, gera possibilidades interessantes na adição de características ao modelo EPE.

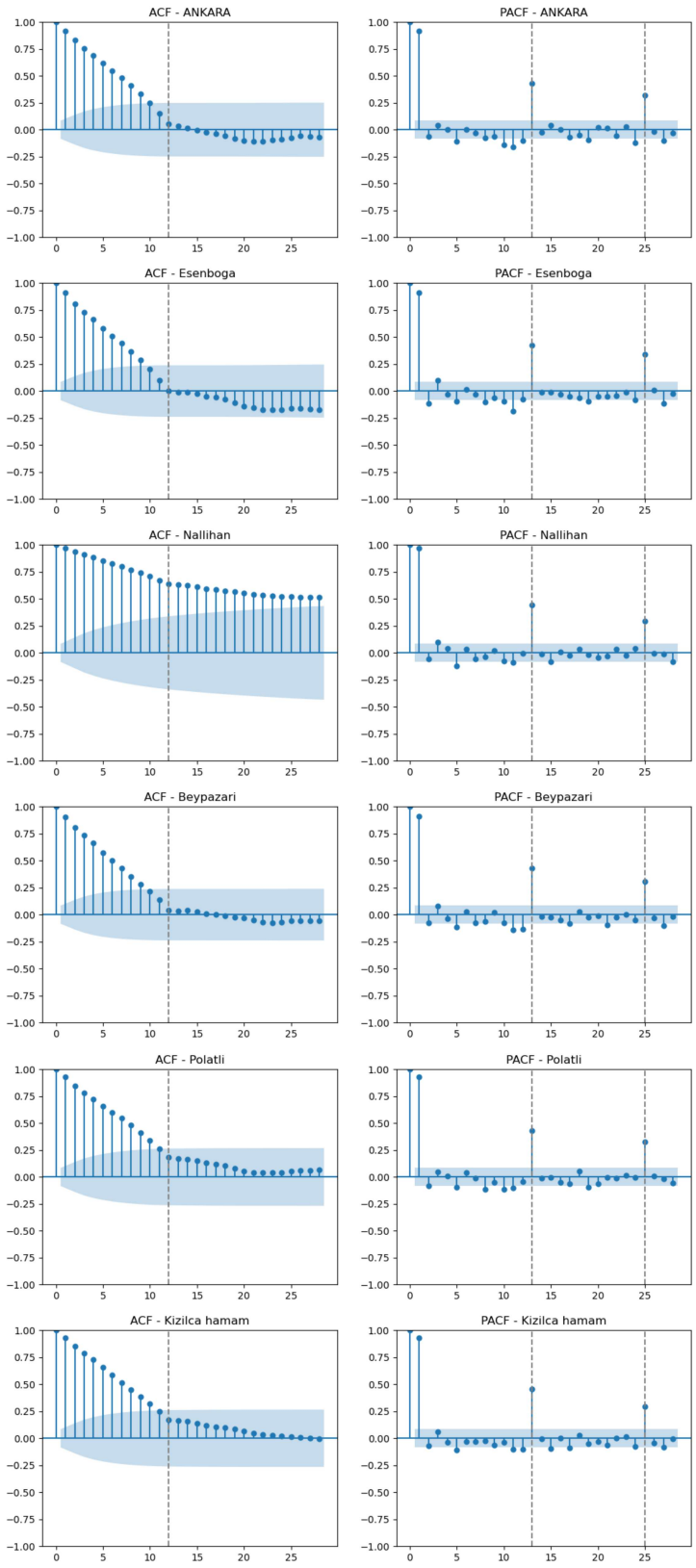


Figura 10 - ACF e PACF para as estações do conjunto de dados.

Além disso, a EDA destaca a importância de mostrar que os eventos de seca se

tornam mais frequentes ao longo do tempo. Esta tendência pode ser visualizada através de gráficos de séries temporais que mostram um aumento na frequência e intensidade dos eventos de seca. Identificar essa tendência é crucial para entender as mudanças climáticas e seus impactos nos recursos hídricos, permitindo que gestores tomem medidas proativas para mitigar os efeitos adversos.

A análise exploratória de dados do SPI, conforme discutido por (97), é vital para fornecer uma compreensão detalhada dos padrões de seca. Isso não só facilita a previsão e gestão de secas, mas também ajuda a desenvolver estratégias de adaptação a longo prazo, essenciais para a sustentabilidade dos recursos hídricos. A utilização de métodos robustos de EDA, como a matriz de correlação, ACF, PACF e gráficos de séries temporais, fornece uma base sólida para decisões informadas e eficazes na gestão de secas

4.2.1 Análise da evolução da amplitude

A análise da evolução da amplitude é uma etapa crítica para entender as mudanças na variabilidade climática ao longo do tempo e como essas mudanças podem influenciar a severidade e a frequência das secas. A amplitude, neste contexto, refere-se a variação entre máximos e mínimos de uma intervalo, o que pode indicar uma maior incerteza ou volatilidade nos padrões de precipitação. Essa análise é fundamental para aprimorar a modelagem preditiva, especialmente em regiões suscetíveis a eventos extremos(98).

A aplicação de técnicas como o Regressor RANSAC (*Random Sample Consensus*) (99) é essencial para detectar outliers e ajustar modelos de regressão robustos que possam captar tendências na evolução do spread ao longo do tempo. O uso do RANSAC é particularmente relevante em dados hidrometeorológicos, onde as séries temporais podem ser influenciadas por anomalias climáticas e ruídos de medição, como demonstrado em trabalhos de (89). A figura 11 ilustra a evolução da ampitude (*spread*) ao longo do tempo para os dados do SPI-12, onde observamos uma tendência de aumento, sugerindo uma maior volatilidade nas condições de seca.

Ao correlacionar o aumento da amplitude com variáveis climáticas, é possível identificar padrões de mudança que são cruciais para a modelagem preditiva. Essa correlação pode ser explorada utilizando técnicas de análise multivariada, como discutido por (100), que permitem compreender melhor as interações entre diferentes variáveis climáticas e o impacto dessas interações na evolução das secas.

Esta análise, portanto, não apenas fornece percepções sobre as tendências de variabilidade climática, mas também aprimora a capacidade dos modelos preditivos de capturar e prever eventos de seca, contribuindo para o desenvolvimento de estratégias eficazes de mitigação e adaptação.

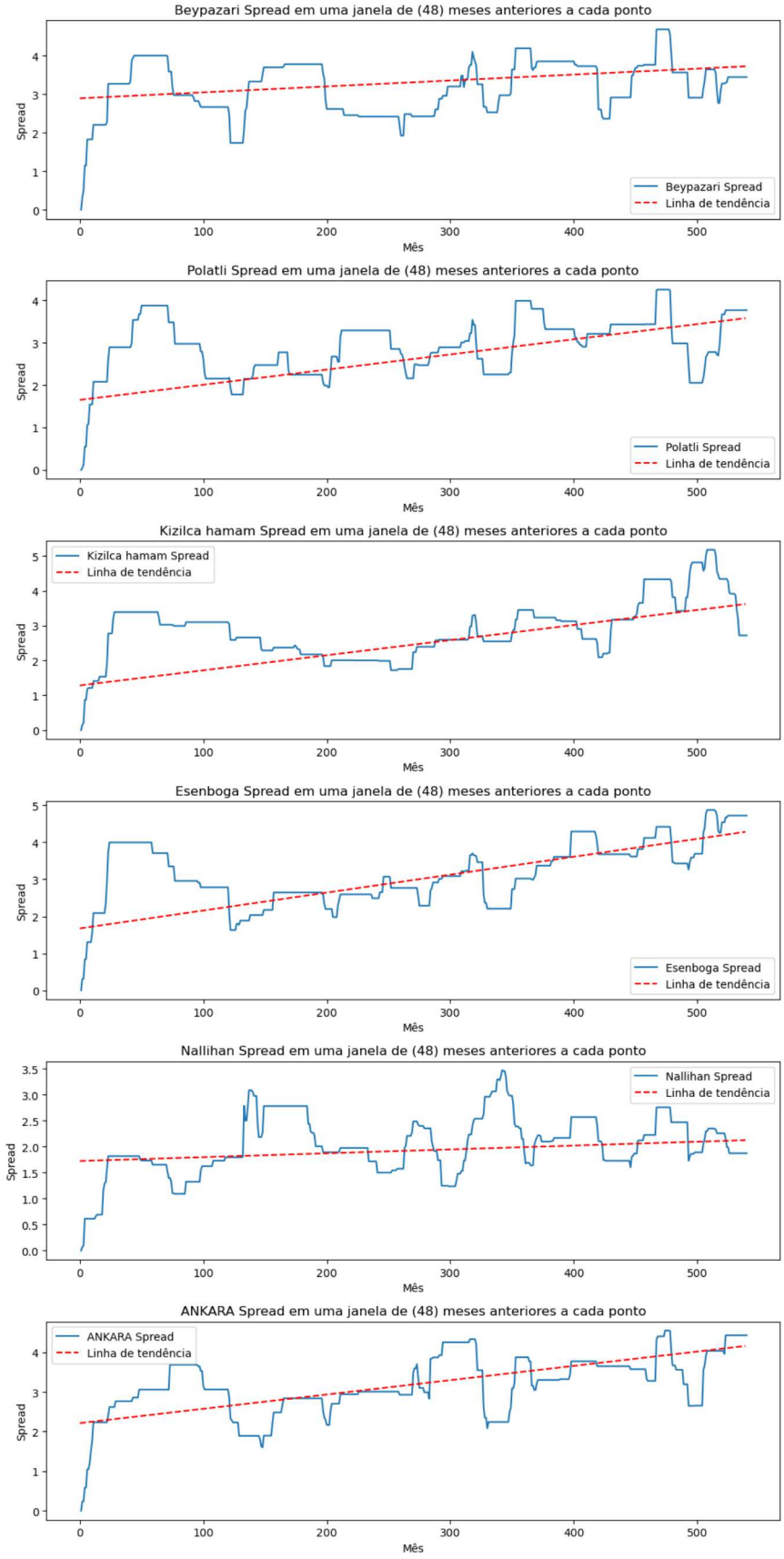


Figura 11 - Análise de evolução da amplitude para os dados de SPI-12.

4.2.2 Importância da Seleção de Características

A seleção de características visa identificar e selecionar o subconjunto mais relevante de variáveis que contribuem significativamente para a previsão do modelo. A escolha adequada de características pode melhorar a precisão do modelo, reduzir a complexidade computacional e aumentar a interpretabilidade dos resultados. Em dados hidrometeorológicos, onde há uma abundância de variáveis potenciais, a seleção de características ajuda a focar nas mais impactantes para a previsão de secas (101).

Dentre as várias técnicas disponíveis para seleção de características, a técnica Lasso (*Least Absolute Shrinkage and Selection Operator*) foi integrada ao modelo EPE. O Lasso é conhecido por sua capacidade de promover a esparsidade ao penalizar a soma absoluta dos coeficientes, resultando em alguns coeficientes exatamente zero e, portanto, eliminando características menos relevantes (102).

O Lasso minimiza a soma residual dos quadrados sujeita à soma do valor absoluto dos coeficientes ser menor que uma constante. Essa restrição induz esparsidade, pois tende a definir alguns coeficientes como exatamente zero, efetivamente selecionando um subconjunto de características que contribuem para o modelo. Isso não só melhora a eficiência do modelo, mas também facilita a interpretabilidade ao reduzir o número de variáveis consideradas.

No contexto do modelo EPE, a técnica Lasso foi utilizada para analisar efetivamente as características dos dados hidrometeorológicos e identificar aquelas que mais contribuem para a previsão do SPI. A integração do Lasso permite que o modelo EPE se concentre na construção de sua estrutura interna utilizando apenas as características mais impactantes, resultando em um modelo eficiente e interpretável. A figura 12 ilustra o funcionamento do modelo EPE.

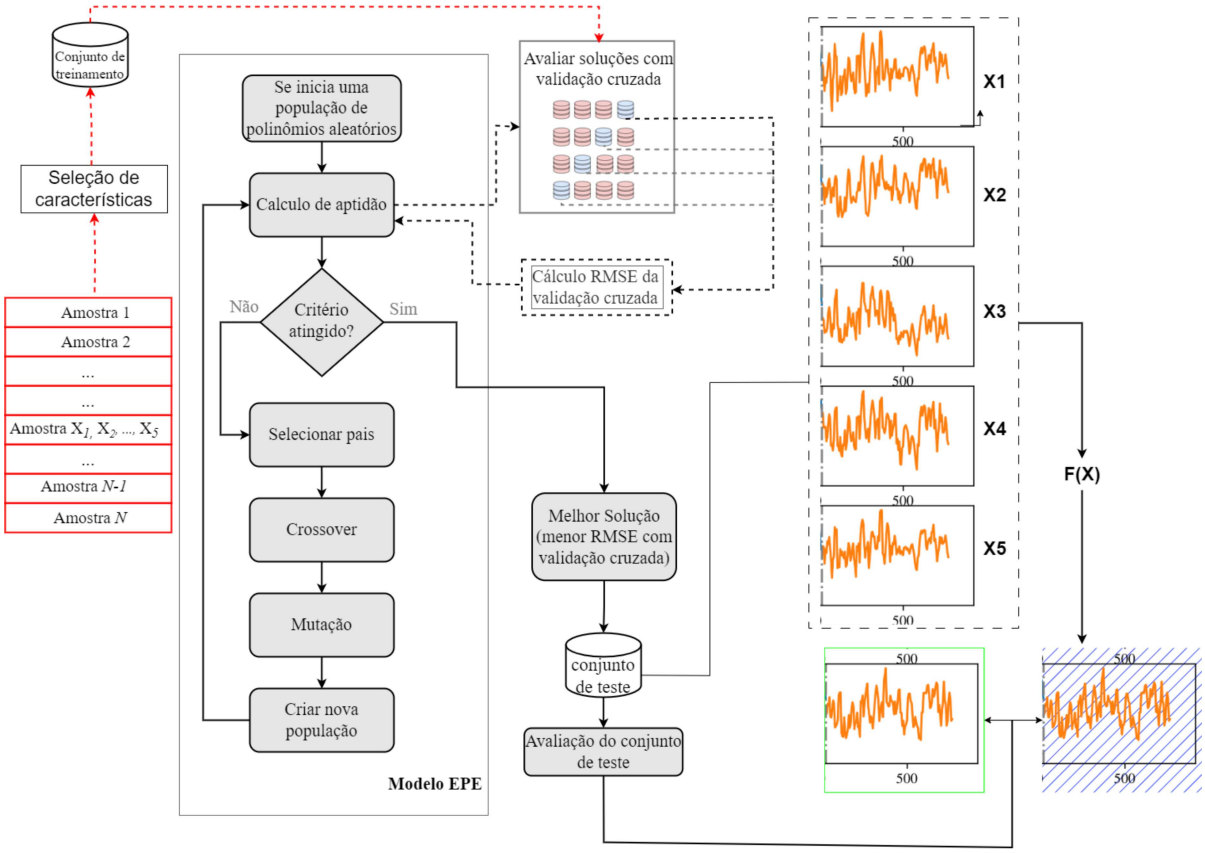


Figura 12 - Funcionamento do modelo EPE.

4.2.3 Inclusão das varíaveis da análise exploratória

Considerando o comportamento sazonal observado e a correlação mostrada anteriormente, as características de *spread* e *lag* foram adicionadas ao modelo EPE em rodadas subsequentes, foram adicionadas:

- *Lag* 12 e 24
- *Spread* 12, 48 e 60

4.2.4 Benchmark com PYSR

O *benchmarking* de modelos preditivos é uma etapa crucial para validar a eficácia de novas abordagens em relação a métodos estabelecidos. Neste estudo, o modelo de Expansão Polinomial Evolutiva (EPE) foi comparado com outras técnicas de regressão simbólica utilizando a ferramenta PYSR (*Python Symbolic Regression*), uma biblioteca moderna que implementa métodos de regressão simbólica altamente eficientes (103). A PYSR se destaca por sua capacidade de encontrar expressões matemáticas que descrevem os dados com alta precisão e simplicidade, utilizando algoritmos baseados em Programação Genética (GP).

A regressão simbólica, como implementada pelo PYSR, oferece uma abordagem interpretável para a modelagem preditiva, uma característica que é fundamental para a compreensão dos mecanismos subjacentes em modelos climáticos e hidrológicos. Trabalhos anteriores, como o de (104), demonstraram a eficácia da Programação Genética na descoberta de modelos simbólicos que capturam complexas interações não lineares entre variáveis. Ao utilizar o PYSR, foi possível obter uma comparação direta entre a EPE e modelos gerados através de GP, destacando as vantagens e limitações de cada abordagem.

A principal métrica utilizada para o benchmarking foi o RMSE (Root Mean Squared Error), uma métrica amplamente aceita para avaliar a precisão preditiva de modelos regressivos. Além disso, foram calculados o coeficiente de determinação (R^2) e a eficiência de Kling-Gupta (KGE) (105), que proporcionam uma visão mais abrangente sobre a qualidade do ajuste dos modelos.

A utilização do PYSR também permite a exploração de diferentes configurações de parâmetros e a análise de sua influência na qualidade do modelo gerado. Essa flexibilidade é essencial para identificar as melhores práticas na aplicação de técnicas de regressão simbólica em contextos climáticos, onde a interpretabilidade dos modelos é tão importante quanto sua precisão (106).

5 RESULTADOS E DISCUSSÃO

A Tabela 5 apresenta as estatísticas descritivas da métrica de desempenho produzida pela metodologia proposta. A primeira coluna mostra o conjunto de dados, a segunda apresenta o β, e as colunas restantes as oito métricas de desempenho. Os resultados mostram que, para SPI-12 e $\beta = 0.5$, as previsões apresentaram melhor desempenho na maioria das métricas, exceto para MAPE.

O β controla a complexidade das expressões polinomiais geradas ao penalizar expressões mais longas. Com $\beta = 0.5$, o modelo alcançou um equilíbrio ideal entre complexidade e precisão, resultando em expressões que eram tanto interpretáveis quanto eficazes na minimização do RMSE. Esse equilíbrio é crucial, pois garante que o modelo permaneça compreensível para os usuários, mantendo um alto desempenho preditivo.

As expressões obtidas para prever secas utilizando a metodologia proposta são apresentadas na Tabela 6. Para cada teste realizado, considerando os conjuntos de dados e os valores de β, foi retornada uma expressão para prever a seca em Ankara (y) usando um conjunto de variáveis que minimiza o RMSE. As expressões que resultaram nos menores valores de RMSE estavam relacionadas ao SPI-12 utilizando 4 variáveis (x_2, x_3, x_4 e x_5) que represantam as outras estações da região.

O modelo destacou a importância de certas variáveis de entrada, especificamente as estações localizadas em Polatlı (x_2), Kızılcahamam (x_3), Esenboğa (x_4) e Nallıhan (x_5), particularmente para o SPI de 12 meses (SPI-12). Essas variáveis foram consistentemente identificadas como preditores críticos, sublinhando sua influência significativa nas condições de seca em Ankara. A inclusão das estações Polatlı, Kızılcahamam, Esenboğa e Nallıhan no modelo EPE ressalta sua importância coletiva em fornecer um conjunto diversificado e representativo de dados climáticos, aumentando a precisão preditiva do modelo para as condições de seca em Ancara. A diversidade geográfica dessas estações captura uma ampla gama de condições climáticas e características geográficas, como planícies, colinas, florestas e áreas urbanas, o que é crucial para compreender as interações complexas e as relações não lineares que afetam os padrões de seca.

Os dados históricos de precipitação, especialmente para o SPI de 12 meses, refletem tendências climáticas de longo prazo e anomalias críticas para a previsão de períodos prolongados de seca, permitindo que o modelo identifique e antecipe as condições de seca com maior precisão. Além disso, estações como Esenboğa, com infraestrutura robusta de coleta de dados, fornecem dados de precipitação de alta qualidade e confiáveis, garantindo que as previsões do modelo sejam baseadas em informações precisas e atualizadas. As localizações estratégicas dessas estações também permitem monitorar os insumos hidrológicos que afetam os recursos hídricos de Ancara, ajudando a avaliar a disponibilidade de água e os impactos potenciais da seca em Ancara de forma abrangente. Esses fatores, coletivamente,

Tabela 5 – Métricas de média e desvio padrão. Desvio padrão entre parênteses.

Dataset	β	R^2	WI	RRMSE	RMSE	MAE	MAPE	NSE	KGE
SPI-3	0.0	0.876 (0.002)	0.966 (0.00)	2.36 (0.020)	0.353 (0.003)	0.273 (0.002)	70.94 (1.50)	0.876 (0.002)	**0.709** (0.088)
	0.1	0.877 (0.002)	**0.967** (0.00)	2.35 (0.015)	0.351 (0.002)	0.272 (0.001)	70.25 (0.808)	0.877 (0.002)	0.693 (0.045)
	0.5	**0.878** (0.00)	0.967 (0.00)	**2.34** (0.004)	**0.350** (0.00)	0.272 (0.001)	70.17 (0.503)	**0.878** (0.00)	0.697 (0.017)
	1.0	0.878 (0.00)	0.967 (0.00)	2.34 (0.007)	0.350 (0.001)	0.272 (0.001)	**70.10** (0.614)	0.878 (0.00)	0.695 (0.019)
	1.5	0.878 (0.00)	0.967 (0.00)	2.34 (0.00)	0.350 (0.00)	**0.271** (0.00)	70.27 (0.436)	0.878 (0.00)	0.693 (0.005)
	2.0	0.878 (0.00)	0.967 (0.00)	2.34 (0.002)	0.350 (0.00)	0.272 (0.001)	70.13 (0.573)	0.878 (0.00)	0.695 (0.008)
PYSR	-	0.878 (0.00)	0.938 (0.00)		0.350 (0.00)				0.728 (0.000)
SPI-6	0.0	0.897 (0.002)	**0.973** (0.00)	**1.70** (0.018)	**0.353** (0.004)	0.292 (0.004)	101.98 (1.56)	0.897 (0.002)	0.809 (0.047)
	0.1	**0.898** (0.006)	0.973 (0.001)	1.70 (0.046)	0.353 (0.010)	0.291 (0.007)	**100.30** (3.45)	**0.898** (0.006)	**0.851** (0.083)
	0.5	0.895 (0.002)	0.972 (0.00)	1.72 (0.015)	0.357 (0.003)	**0.286** (0.002)	109.67 (2.60)	0.895 (0.002)	0.662 (0.048)
	1.0	0.895 (0.00)	0.972 (0.00)	1.72 (0.001)	0.357 (0.00)	0.286 (0.00)	110.15 (0.389)	0.895 (0.00)	0.654 (0.002)
	1.5	0.895 (0.001)	0.972 (0.00)	1.72 (0.011)	0.357 (0.002)	0.286 (0.00)	109.64 (2.50)	0.895 (0.001)	0.663 (0.048)
	2.0	0.895 (0.001)	0.972 (0.00)	1.72 (0.010)	0.357 (0.002)	0.286 (0.00)	109.52 (2.46)	0.895 (0.001)	0.664 (0.046)
PYSR	-	0.898 (0.000)			0.352 (0.00)				0.698 (0.000)
SPI-12	0.0	0.902 (0.003)	0.972 (0.001)	1.22 (0.022)	0.349 (0.006)	0.285 (0.005)	123.01 (8.32)	0.902 (0.003)	0.712 (0.067)
	0.1	0.907 (0.005)	0.974 (0.002)	1.19 (0.032)	0.340 (0.009)	0.278 (0.007)	117.01 (5.16)	0.907 (0.005)	0.774 (0.103)
	0.5	**0.914** (0.00)	**0.977** (0.00)	**1.14** (0.002)	**0.326** (0.00)	**0.268** (0.00)	113.14 (1.01)	**0.914** (0.00)	**0.911** (0.012)
	1.0	0.913 (0.003)	0.977 (0.00)	1.15 (0.018)	0.328 (0.005)	0.269 (0.003)	112.01 (5.49)	0.913 (0.003)	0.892 (0.060)
	1.5	0.913 (0.004)	0.977 (0.001)	1.15 (0.025)	0.329 (0.007)	0.269 (0.005)	111.30 (6.51)	0.913 (0.004)	0.881 (0.077)
	2.0	0.911 (0.006)	0.977 (0.001)	1.16 (0.038)	0.332 (0.011)	0.271 (0.006)	**106.86** (11.45)	0.911 (0.006)	0.845 (0.117)
PYSR	-	0.910 (0.000)			0.333 (0.000)				0.674 (0.000)

destacam a importância dessas estações na melhoria da capacidade do modelo em prever com precisão as condições de seca.

Tabela 6 – Expressões para previsão de seca.

Dataset	β	# var.	RMSE	Expression (y)
SPI-3	0.0	5	0.348	$x_1 \cdot (0.065438x_4 - 0.031779x_5) + x_2 \cdot (0.4524 - 0.038433x_3) + x_3 \cdot (0.059723 - 0.016967x_4) + 0.498031x_4 - 0.049476x_5 + 0.006301$
	0.1	3	0.348	$0.455642x_2 + x_4 \cdot (0.501748 - 0.006454x_3) + 0.021436$
	0.5	3	**0.347**	$0.449359x_2 + 0.057424x_3 + 0.475391x_4 + 0.009358$
	1.0	4	**0.347**	$0.447717x_2 + x_3 \cdot \left(0.056808 - \frac{0.001706}{x_5}\right) + 0.476609x_4 + 0.009888$
	1.5	3	0.349	$0.454934x_2 + x_4 \cdot \left(0.502252 - \frac{0.001708}{x_5}\right) + 0.018129$
	2.0	2	0.349	$0.451833x_2 + 0.495976x_4 + 0.016304$
SPI-6	0.0	5	0.349	$x_1 \cdot (0.045872x_2 + 0.020121x_4 - 0.038107x_5) + x_2 \left(-0.040114x_3 + 0.490713 - \frac{0.003131}{x_3}\right) + x_4 \cdot \left(0.545761 + \frac{0.003146}{x_5}\right) + x_5 \left(-0.095141 + \frac{0.000671}{x_3}\right) + 0.003401$
	0.1	5	0.346	$x_2 \cdot (0.490102 - 0.017279x_3) + x_4 \cdot \left(0.018656x_1 + 0.540545 + \frac{0.002757}{x_5}\right) - 0.09504x_5 + 0.001836$
	0.5	3	0.346	$0.490587x_2 + 0.55039x_4 - 0.107172x_5 + 0.003939$
	1.0	3	0.357	$0.481616x_2 + x_4 \cdot \left(0.50263 + \frac{0.003083}{x_5}\right) + 0.028295$
	1.5	4	**0.345**	$x_2 \cdot (0.487884 - 0.003141x_3) + 0.546381x_4 - 0.10162x_5 + 0.006395$
	2.0	3	0.346	$0.490585x_2 + 0.550387x_4 - 0.107168x_5 + 0.003939$
SPI-12	0.0	5	0.328	$x_1 \cdot (0.051353x_4 - 0.038817x_5) + 0.464196x_2 + 0.604415x_4 + x_5 (-0.062861x_3 - 0.153804) - 0.018482 + \frac{0.001541x_4 - 0.001864x_5}{x_1}$
	0.1	4	0.324	$0.470079x_2 - \frac{0.000222x_3}{x_4} + 0.581248x_4 - 0.141698x_5 - 0.001256$
	0.5	3	0.325	$0.476589x_2 + 0.588741x_4 - 0.15268x_5 - 0.003222$
	1.0	3	0.325	$0.47658x_2 + 0.588731x_4 - 0.152665x_5 - 0.00322$
	1.5	3	0.325	$0.475971x_2 + 0.587893x_4 - 0.151498x_5 - 0.00305$
	2.0	4	**0.319**	$0.466306x_2 + 0.590777x_4 + x_5 (-0.073891x_3 - 0.139466) + 0.009325$

x_1(Beypazari), x_2(Polatli), x_3(Kizilca Hamam), x_4(Esenboga), x_5(Nallihan), x_6(Kecioren), y(Ankara).

A Figura 13 ilustra a relação entre β, o número de variáveis selecionadas, e o Erro Quadrático Médio (RMSE) para cada conjunto de dados (SPI-3, SPI-6, SPI-12). Os gráficos revelam uma tendência clara: um valor menor de β leva à seleção de um maior número de variáveis. Isso é evidente no número de variáveis para todos os conjuntos de dados.

A Figura 14 mostra a frequência de seleção das variáveis. Pode-se observar que o

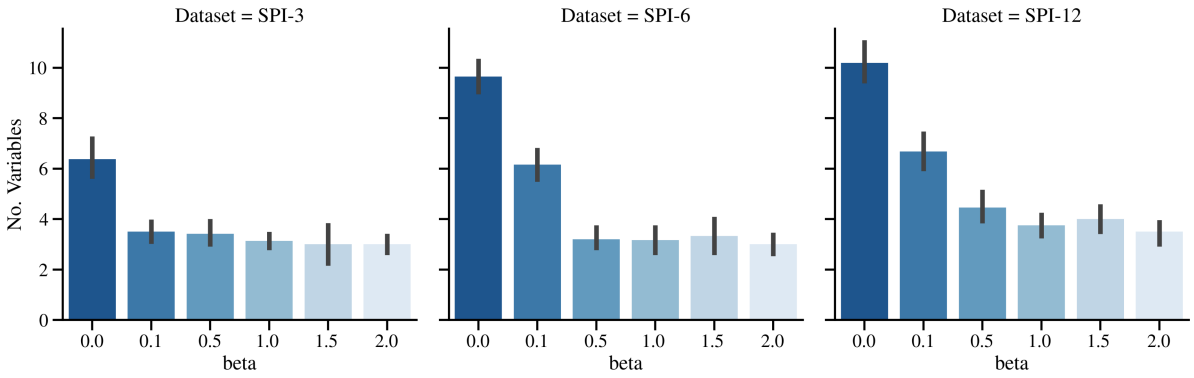


Figura 13 - Distribuição do número de variáveis

SPI-3 e o SPI-6 exibem comportamentos semelhantes. Em ambos os conjuntos de dados, modelos contendo três variáveis são frequentemente escolhidos em diferentes configurações de modelos. Isso sugere que modelos com três variáveis podem ter um poder preditivo significativo para a seca nesses conjuntos de dados. No entanto, o SPI-12 apresenta um padrão distinto, com modelos de quatro variáveis sendo os mais frequentemente selecionados.

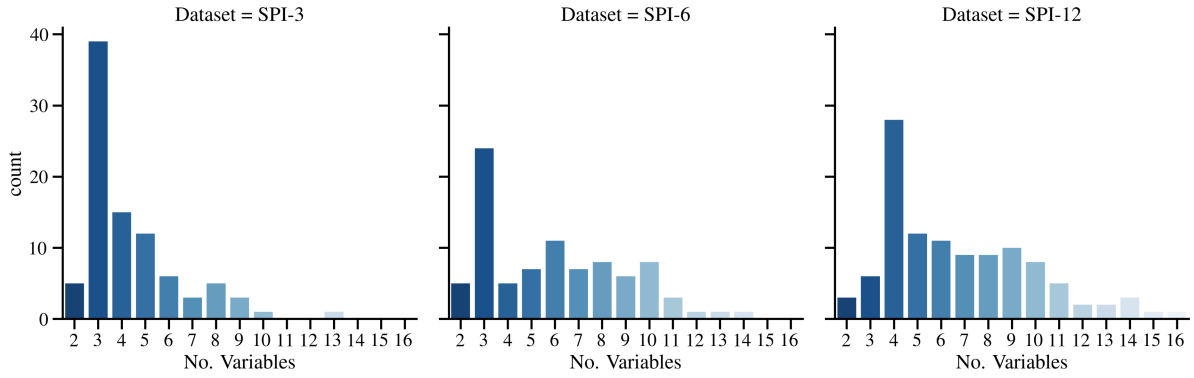


Figura 14 - Numero de variáveis de acordo com o parâmetro β.

A Figura 15 apresenta as estatísticas de beta e o número de variáveis nos conjuntos de dados. Os gráficos de número de variáveis por beta mostram que, para todos os conjuntos de dados, o valor de β influencia a seleção de variáveis; para um valor de β mais baixo, mais variáveis são selecionadas, enquanto para um valor de β mais alto, o número de variáveis selecionadas é menor. Nos gráficos de contagem por número de variáveis, para SPI-3, a variável selecionada com mais frequência foi a variável 3; para SPI-6, a variável 3 também foi a mais selecionada, mostrando que teve grande influência nos resultados desses conjuntos de dados. No entanto, para o conjunto de dados SPI-12, a variável 4 foi a mais selecionada. Finalmente, nos gráficos de RMSE por número de variáveis, é evidente que, para o conjunto de dados SPI-3, os menores valores de RMSE ocorrem para um maior número de variáveis, o que também acontece para o conjunto de dados SPI-6. No entanto, para o conjunto de dados SPI-12, os menores valores de RMSE ocorrem para um menor

número de variáveis, mostrando que, para este conjunto de dados, um número menor de variáveis influencia a previsão de seca.

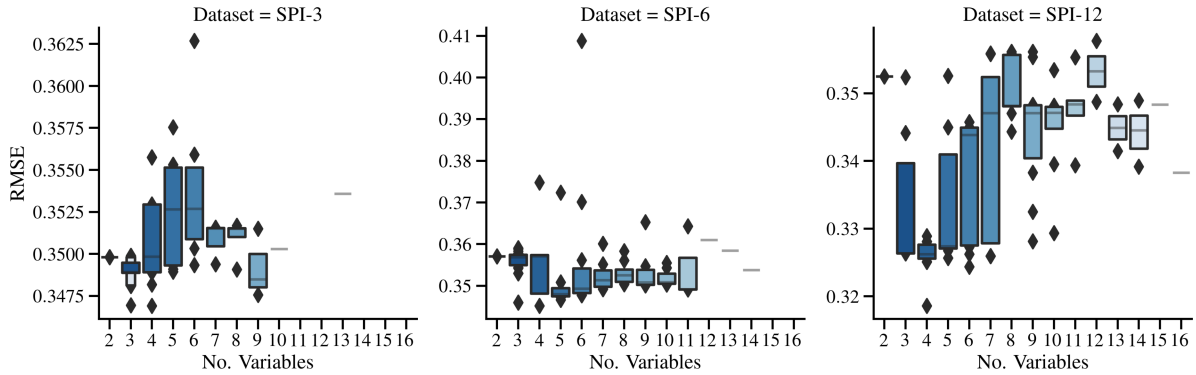


Figura 15 - Boxplots para RMSE de acordo com o numero de variáveis.

A Figura 16 apresenta diagramas de dispersão que mostram o desempenho dos modelos de regressão Lasso otimizados com diferentes valores de β para cada estação nos conjuntos de teste. Os resultados indicam que o conjunto de dados SPI-12 alcançou os maiores valores de R-quadrado, sugerindo o melhor ajuste geral à linha de regressão. Dentro do conjunto SPI-12, os modelos com β igual a 0.1 e 0.5 exibiram o melhor desempenho com base nas métricas de avaliação. Essa observação destaca a influência de β no desempenho do modelo, com valores mais altos de β geralmente levando a melhores pontuações nas métricas.

Os modelos de aprendizado de máquina interpretáveis ganharam considerável tração nos últimos anos segundo (107, 108), impulsionados pela necessidade de entender e explicar as previsões dos modelos. Expressões interpretáveis, particularmente aquelas derivadas da regressão simbólica, oferecem uma ferramenta valiosa para decifrar o comportamento do modelo e obter insights sobre as relações subjacentes entre as características de entrada e a saída. No entanto, formular expressões interpretáveis enquanto se mantém o desempenho do modelo continua sendo um desafio formidável.

Nesta dissertação, introduzimos um parâmetro de controle de complexidade, denotado como β, para regular a complexidade das expressões geradas. Este parâmetro é incorporado na função objetivo, conforme mostrado na Equação (3.7), influenciando indiretamente a complexidade das expressões. Embora o parâmetro de controle de complexidade ofereça uma abordagem simples para regular a complexidade das expressões, apresenta certas limitações. Primeiramente, ele depende da seleção de parâmetros definidos pelo usuário, que pode nem sempre estar alinhada com o nível ideal de complexidade para um dado problema. Em segundo lugar, a relação entre β e a complexidade das expressões pode não ser totalmente capturada, potencialmente levando a resultados sub ótimos.

Uma abordagem alternativa para lidar com a troca entre complexidade e interpretabilidade envolve o emprego de técnicas de otimização multiobjetivo. Essa estratégia

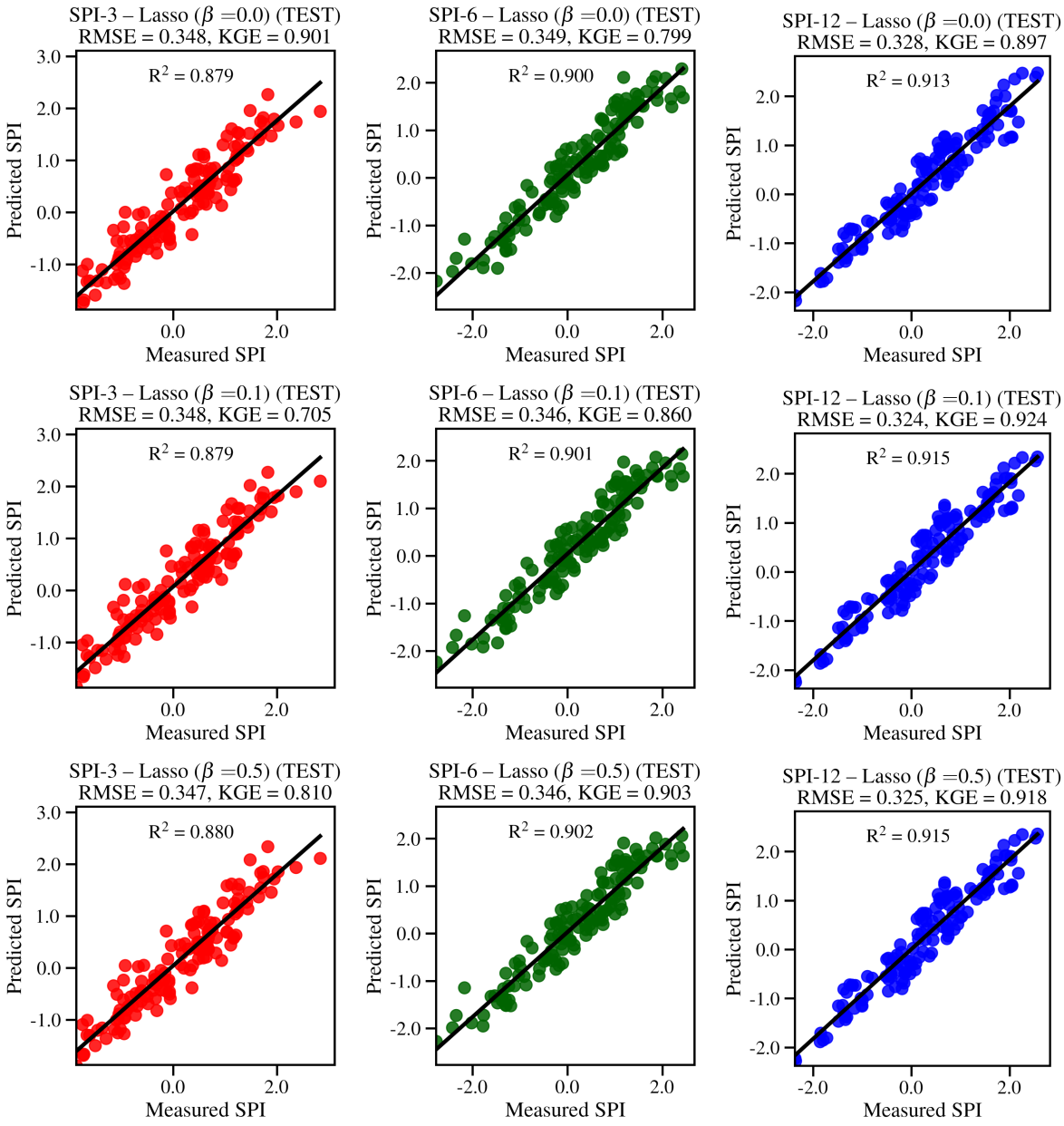


Figura 16 - Dispersão dos melhores modelos por β e dataset.

consiste em definir objetivos conflitantes para o problema, como minimizar tanto a complexidade das expressões quanto o erro de previsão, criando uma estratégia alternativa para equilibrar complexidade e interpretabilidade. Essa abordagem assume que expressões mais simples geralmente apresentam menor precisão preditiva. Algoritmos de otimização multiobjetivo visam identificar soluções não dominadas ao longo da fronteira de Pareto. Esse processo garante que as melhores soluções possíveis sejam obtidas, mitigando o problema de expressões com precisão semelhante, mas complexidades variáveis.

Pesquisas futuras envolvem a formulação multiobjetivo (85) para modelar a troca entre complexidade e precisão, gerando uma população de modelos que servem como soluções alternativas enquanto auxiliam na tomada de decisão.

6 CONCLUSÃO

A presente dissertação investigou a aplicação de modelos de aprendizado de máquina na previsão do Índice Padronizado de Precipitação (SPI), com foco especial na complexidade e interpretabilidade das expressões geradas. Os resultados obtidos demonstram que a utilização do modelo de Expansão Polinomial Evolutiva (EPE), aliado à seleção de características Lasso, oferece uma abordagem robusta para a previsão de secas, equilibrando precisão e interpretabilidade.

Os principais achados deste estudo indicam que o controle de complexidade, através do parâmetro $\beta = 0$, encontrado através dos experimentos, é essencial para gerar modelos que são ao mesmo tempo eficazes e compreensíveis. A integração dos parâmetros dentro da função objetivo, mostrou-se eficaz na identificação de soluções que atendem a ambos os critérios, oferecendo um conjunto de modelos alternativos que podem ser utilizados conforme a necessidade específica da aplicação.

Além disso, a análise exploratória dos dados, utilizando técnicas como a função de autocorrelação (ACF), função de autocorrelação parcial (PACF) e a matriz de correlação, proporcionou uma compreensão detalhada dos padrões de seca, essenciais para a seleção de características relevantes se sazonalidade anual (12 e 24 meses) e para a construção de modelos preditivos mais precisos. Além de indicar, como sugere a literatura, o aumento dos eventos de seca ao longo do tempo.

A validação cruzada de séries temporais provou ser uma abordagem robusta para avaliar o desempenho dos modelos, evitando o vazamento de informações e garantindo estimativas imparciais. As métricas de desempenho encotradas, $KGE = (0.709,\ 0.851,\ 0.911)$, para SPI-3, 6 e 12 respectivamente e , $R^2 = (0.878,\ 0.898,\ 0.977)$ e $RMSE = (0.350,\ 0.353,\ 0.326)$, permitiram uma avaliação abrangente da eficácia dos modelos gerados, destacando a importância de equilibrar a precisão com a complexidade das expressões. Essas métricas foram balizadas pela solução de mercado PYSR e mostraram estar em níveis de desempenho similares.

Em suma, a pesquisa contribui para o campo da previsão de secas, oferecendo uma metodologia que não apenas melhora a precisão das previsões, mas também mantém a interpretabilidade das expressões, facilitando a tomada de decisões informadas por parte de gestores de recursos hídricos e especialistas em clima. Pesquisas futuras podem explorar a integração de outras técnicas de aprendizado de máquina e a aplicação do modelo proposto em diferentes contextos climáticos, ampliando ainda mais a aplicabilidade e a eficácia dos métodos desenvolvidos.

6.1 TRABALHOS FUTUROS

A fim de aprimorar a eficácia e a aplicabilidade da modelagem de secas utilizando o modelo de Expansão Polinomial Evolutiva (EPE), diversas linhas de investigação e desenvolvimento podem ser exploradas:

- **Expansão do Modelo para Previsão em t+1:** Uma direção promissora para futuras pesquisas é a expansão do modelo EPE para realizar previsões em um horizonte temporal futuro, como t+1. Este avanço permitiria não apenas a detecção de eventos de seca em andamento, mas também a antecipação de futuras anomalias hidrometeorológicas. Modelos de previsão em séries temporais, como aqueles descritos por (109) e (110), demonstram o potencial de abordagens que integram aprendizado de máquina com modelagem de séries temporais para melhorar a precisão de previsões climáticas.

- **Aprimoramento da Inclusão de Variáveis da EDA:** A inclusão das variáveis selecionadas durante a Análise Exploratória de Dados (EDA) no modelo EPE deve ser otimizada para assegurar que características críticas sejam eficazmente capturadas e que o modelo mantenha sua interpretabilidade. Métodos avançados de seleção de características, como os propostos por (111), podem ser integrados ao processo de EDA para identificar as variáveis mais influentes de maneira mais robusta, melhorando assim a qualidade das previsões.

- **Desenvolvimento de um Sistema de Monitoramento em Tempo Real:** A transição do modelo EPE para um sistema de monitoramento de secas em tempo real representa um passo crucial para a aplicação prática deste estudo. Este sistema pode ser desenvolvido para fornecer alertas precoces e insights acionáveis para a gestão de recursos hídricos e mitigação de desastres naturais. Estudos de caso como o de (112) enfatizam a importância de sistemas de monitoramento em tempo real para a gestão proativa de secas.

- **Aprimoramento da Definição do Parâmetro Beta:** O parâmetro β no modelo EPE desempenha um papel crítico ao balancear a complexidade e a precisão do modelo. Pesquisas futuras devem focar no desenvolvimento de métodos mais sofisticados para a definição e ajuste deste parâmetro, possivelmente utilizando técnicas de otimização evolutiva ou algoritmos de busca global, como discutido por (113). A definição adequada de β é fundamental para assegurar que o modelo gere expressões polinomiais que sejam ao mesmo tempo precisas e interpretáveis.

- **Inclusão de Variáveis Geográficas na Análise:** A inclusão de variáveis geográficas, como elevação, uso do solo e características geomorfológicas, pode enriquecer a análise, permitindo ao modelo EPE capturar de maneira mais completa as interações

espaciais que influenciam os eventos de seca. Abordagens integradas que combinam dados climáticos com informações geográficas, como explorado por (114), podem revelar padrões espaciais críticos que influenciam a severidade e a extensão das secas, oferecendo uma visão mais holística do fenômeno.

Essas sugestões delineiam um caminho claro para avanços significativos no campo da modelagem de secas, alavancando as capacidades do modelo EPE e promovendo sua aplicação prática em contextos reais de monitoramento e gestão de secas.

REFERÊNCIAS

1 Benjamin I. Cook, Justin S. Mankin, and Kevin J. Anchukaitis. Climate change and drought: From past to future. *Current Climate Change Reports*, 4(2):164–179, May 2018. doi: 10.1007/s40641-018-0093-2.

2 S. L. Shah, B. R. Bakshi, J. Liu, C. Georgakis, B. Chachuat, R. D. Braatz, and B. R. Young. Meeting the challenge of water sustainability: The role of process systems engineering. *AIChE Journal*, 2020. doi: 10.1002/aic.17113.

3 E. E. Ebhuoma, F. K. Donkor, O. O. Ebhuoma, L. Leonard, and H. B. Tantoh. Subsistence farmers' differential vulnerability to drought in mpumalanga province, south africa: Under the political ecology spotlight. *Cogent Social Sciences*, 6(1): 1792155, 2020. doi: 10.1080/23311886.2020.1792155.

4 Hesami Afshar. Mahdi and sorman, ali unal and yilmaz, mustafa tugrul (2016). conditional copula-based spatial–temporal drought characteristics analysis—a case study over turkey. *Water*, 8(10):426.

5 Thomas B. McKee, Nolan J. Doesken, and John Kleist. The relationship of drought frequency and duration to time scales. In *Eighth Conference on Applied Climatology*, pages 179–183, American Meteorological Society, 1993. 17(22). URL `https://www.droughtmanagement.info/literature/AMS_Relationship_Drought_Frequency_Duration_Time_Scales_1993.pdf`.

6 T. H. Li and W. C. Xie. A new method for computing the sediment delivery ratio for the hyper-concentrated flow areas of the loess plateau, China. *Journal of Environmental Informatics*, 39(1):1–10, 2022.

7 Amin Zargar, Rehan Sadiq, Bahman Naser, and Faisal I. Khan. A review of drought indices. *Environmental Reviews*, 19(NA), . NRC Research Press:333–349, 2011.

8 Hi-Ryong Byun and Donald A. Wilhite. Objective quantification of drought severity and duration. *Journal of Climate*, 12(9):2747–2756, 1999.

9 Singh T. P. Nandgude S. Nandgude, N. and M. Tiwari. Drought prediction: A comprehensive review of different drought prediction models and adopted technologies. *Sustainability*, 15(15):11684, July 2023. doi: 10.3390/su151511684. URL `https://www.mdpi.com/2071-1050/15/15/11684`.

10 Majed Alsubih, Javed Mallick, Swapan Talukdar, Roquia Salam, Saeed AlQadhi, and Md Fattah. Abdul and thanh, nguyen viet (2021). an investigation of the short-term meteorological drought variability over asir region of saudi arabia. *Theoretical and Applied Climatology*, 145(1-2):597–617. doi: 10.1007/s00704-021-03647-4.

11 A.A. Asr, A. Johari, and A.A. Javadi. An evolutionary-based polynomial regression modeling approach to predicting discharge flow rate under sheet piles. *Engineering with Computers*, 2023. URL `https://link.springer.com/article/10.1007/s00366-023-01872-1`. Accessed: 21 August 2024.

12 Renata Sadrtdinova, Gerald Augusto Corzo Perez, and Dimitri P. Solomatine. Improved drought forecasting in kazakhstan using machine and deep learning: a non-contiguous drought analysis approach. *Hydrology Research*, 55(2):237–261, 2024. doi: 10.2166/nh.2024.154.

13 Sanaa Hobeichi, Gab Abramowitz, and Jason P. Evans. and ukkola, anna (2022). toward a robust, impact-based, predictive drought metric. *Water Resources Research*, 58(2). doi: 10.1029/2021WR031829.

14 Mohammed Achite, Nehal Elshaboury, Muhammad Jehanzaib, Dinesh Kumar Vishwakarma, Quoc Bao Pham, Duong Tran Anh, Eslam Mohammed Abdelkader, and Ahmed Elbeltagi. Performance of machine learning techniques for meteorological drought forecasting in the wadi mina basin, algeria. *Water (Switzerland)*, 15(4), 2023. doi: 10.3390/w15040765.

15 Nu Htway. Nu and thin, lwin mar (2023). comparison of machine learning models for drought prediction in central myanmar. In *2023 International Conference on the Confluence of Advancements in Robotics, Vision and Interdisciplinary Technology Management, IC-RVITM 2023*. doi: 10.1109/IC-RVITM60032.2023.10435106.

16 P. Sandhya Krishna and B. Yamini Krishna. and nafisa. In Shaik, editor, *Proceedings - 2023 12th IEEE International Conference on Communication Systems and Network Technologies, CSNT 2023*, pages 839–845, 2023. doi: 10.1109/CSNT57126.2023.10134742.

17 Rodrigue B. W. Vodounon, Henoc Soude, and Ossénatou Mamadou. Drought forecasting in alibori department in benin using the standardized precipitation index and machine learning approaches. *International Journal of Advanced Computer Science and Applications*, 13(12):987–994, 2022. doi: 10.14569/IJACSA.2022.01312113.

18 Heri Kuswanto and Achmad Naufal. Evaluation of performance of drought prediction in indonesia based on trmm and merra-2 using machine learning methods. *MethodsX*, 6():1238–1251, 2019. doi: 10.1016/j.mex.2019.05.029.

19 Ahmed Elbeltagi and Chaitanya B. Pande. and kumar, manish and tolche, abebe debele and singh, sudhir kumar and kumar, akshay and vishwakarma, dinesh kumar (2023). prediction of meteorological drought and standardized precipitation index based on the random forest (rf), random tree (rt), and Gaussian process regression (gpr) models. *Environmental Science and Pollution Research*, 30(15):43183–43202. doi: 10.1007/s11356-023-25221-3.

20 Sushree Swagatika Swain, Ashok Mishra, and Chandranath Chatterjee. Time-varying evaluation of compound drought and hot extremes in machine learning-predicted ensemble cmip5 future climate: A multivariate multi-index approach. *Journal of Hydrologic Engineering*, 29(2), 2024. doi: 10.1061/JHYEFF.HEENG-6026.

21 Hannah Kemper. Development of a drought early warning system based on the prediction of agricultural productivity: A data science approach. *GI_Forum*, 10(1): 58–76, 2022. doi: 10.1553/giscience2022-01-s58.

22 Puyu Feng, Bin Wang, Jing-Jia Luo, De Li Liu, Cathy Waters, Fei Ji, Hongyan Ruan, Dengpan Xiao, Lijie Shi, and Qiang Yu. Using large-scale climate drivers to forecast meteorological drought condition in growing season across the australian wheatbelt. *Science of the Total Environment*, 724(), 2020. doi: 10.1016/j.scitotenv.2020.138162.

23 Safwan Mohammed, Ahmed Elbeltagi, Bashar Bashir, Karam Alsafadi, Firas Alsilibe, Abdullah Alsalman, Mojtaba Zeraatpisheh, Adrienn Széles, and Endre Harsányi. A comparative analysis of data mining techniques for agricultural and hydrological drought prediction in the eastern mediterranean. *Computers and Electronics in Agriculture*, 197(), 2022. doi: 10.1016/j.compag.2022.106925.

24 Md Sadiq. Ashhab and sarkar, showmitra kumar and raisa, saima sekander (2023). meteorological drought assessment in northern bangladesh: A machine learning-based approach considering remote sensing indices. *Ecological Indicators*, 157(). doi: 10.1016/j.ecolind.2023.111233.

25 Mahmood Fooladi, Mohammad H. Golmohammadi, Hamid R. Safavi, and Vijay P. Singh. Fusion-based framework for meteorological drought modeling using remotely sensed datasets under climate change scenarios: Resilience, vulnerability, and frequency analysis. *Journal of Environmental Management*, 297(), 2021. doi: 10.1016/j.jenvman.2021.113283.

26 Samuel J. Sutanto and Van Der Weert. Melati and blauhut, veit and van lanen, henny a. *J. Skill of large-scale seasonal drought impact forecasts.*, 20(6):1595–1608, 2020. doi: 10.5194/nhess-20-1595-2020.

27 Morteza Lotfirad, Hassan Esmaeili-Gisavandani, and Arash Adib. Drought monitoring and prediction using spi, spei, and random forest model in various climates of iran. *Journal of Water and Climate Change*, 13(2), 2022. doi: 10.2166/wcc.2021.287.

28 Roghayeh Ghasempour, Mohammad Taghi Aalami, V. S. Ozgur Kirca, and Kiyoumars Roushangar. Remote sensing-based drought severity modeling and mapping using multiscale intelligence methods. *Stochastic Environmental Research and Risk Assessment*, 37(3):889–902, 2023. doi: 10.1007/s00477-022-02324-w.

29 Mei-Yan Zhao, Tao Hu, Yu-Hu Zhang, Xiao Pu, and Feng Gao. Drought prediction based on machine learning models in the northern part of haihe river basin. *Arid Land Geography*, 43(4):880–888, 2020. doi: 10.12118/j.issn.1000-6060.2020.04.03.

30 A. J. Gibson, D. C. Verdon-Kidd, and G. R. Hancock. Characterising the seasonal nature of meteorological drought onset and termination across australia. *Journal of Southern Hemisphere Earth Systems Science*, 72(1):38–51, 2022. doi: 10.1071/ES21009.

31 M. A. Jincy Rose and N. R. Chithra. Tree-based ensemble model prediction for hydrological drought in a tropical river basin of india. *International Journal of Environmental Science and Technology*, 20(5):4973–4990, 2023. doi: 10.1007/s13762-022-04208-6.

32 Abdol Rassoul Zarei, Mohammad Reza Mahmoudi, and Mohammad Mehdi Moghimi. Determining the most appropriate drought index using the random forest algorithm

with an emphasis on agricultural drought. *Natural Hazards*, 115(1):923–946, 2023. doi: 10.1007/s11069-022-05579-2.

33 Soo-Jin Lee, Eunha Sohn, Mija Kim, Ki-Hong Park, Kyungwon Park, and Yangwon Lee. Real-time retrieval of daily soil moisture using imerg and gk2a satellite images with nwp and topographic data: A machine learning approach for south korea. *Remote Sensing*, 15(17), 2023. doi: 10.3390/rs15174168.

34 Zaher Mundher Yaseen, Mumtaz Ali, Ahmad Sharafati, Nadhir Al-Ansari, and Shamsuddin Shahid. Forecasting standardized precipitation index using data intelligence models: regional investigation of bangladesh. *Scientific Reports*, 11(1), 2021. doi: 10.1038/s41598-021-82977-9.

35 Ziyue Li, Zhao Zhang, and Lingyan Zhang. Improving regional wheat drought risk assessment for insurance application by integrating scenario-driven crop model, machine learning, and satellite data. *Agricultural Systems*, 191(), 2021. doi: 10.1016/j.agsy.2021.103141.

36 Safwan Mohammed, Sana Arshad, Firas Alsilibe, Muhammad Farhan Ul Moazzam, Bashar Bashir, Foyez Ahmed Prodhan, Abdullah Alsalman, Attila Vad, Tamás Ratonyi, and Endre Harsányi. Utilizing machine learning and cmip6 projections for short-term agricultural drought monitoring in central europe (1900–2100). *Journal of Hydrology*, 633, 2024. doi: 10.1016/j.jhydrol.2024.130968. URL `https://www.scopus.com/inward/record.uri?eid=2-s2.0-85187128031&doi=10.1016%2fj.jhydrol.2024.130968&partnerID=40&md5=fd817479732fa979398b8eca55506dde`. All Open Access, Hybrid Gold Open Access.

37 J. Rhee, J. Im, and S. Park. Drought forecasting based on machine learning of remote sensing and long-range forecast data. *International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences - ISPRS Archives*, 41():157–158, 2016. doi: 10.5194/isprsarchives-XLI-B8-157-2016.

38 Seonyoung Park, Jungho Im, Eunna Jang, and Jinyoung Rhee. Drought assessment and monitoring through blending of multi-sensor indices using machine learning approaches for different climate regions. *Agricultural and Forest Meteorology*, 216(): 157–169, 2016. doi: 10.1016/j.agrformet.2015.10.011.

39 A. Jalalkamali, M. Moradi, and N. Moradi. Application of several artificial intelligence models and arimax model for forecasting drought using the standardized precipitation index. *International Journal of Environmental Science and Technology*, 12(4):1201–1210, 2015. doi: 10.1007/s13762-014-0717-6.

40 Lilik Budi Prasetyo, Yudi Setiawan, Aryo Adhi Condro, Kustiyo Kustiyo, Erianto Indra Putra, Nur Hayati, Arif Kurnia Wijayanto, Almi Ramadhi, and Daniel Murdiyarso. Assessing sumatran peat vulnerability to fire under various condition of enso phases using machine learning approaches. *Forests*, 13(6), 2022. doi: 10.3390/f13060828.

41 Ali Danandeh Mehr. Drought classification using gradient boosting decision tree. *Acta Geophysica*, 69(3):909–918, 2021. doi: 10.1007/s11600-021-00584-8.

42 Poulomi Ganguli and Janga Reddy. M. *Ensemble prediction of regional droughts using climate inputs and the SVM-copula approach.*, 28(19):4989–5009, 2014. doi: 10.1002/hyp.9966.

43 Ali Reza Nikbakht Shahbazi, Banafsheh Zahraie, Hossein Sedghi, and Mohammad and Manshouri. and.

44 Hatice Citakoglu and Omer Coşkun. Comparison of hybrid machine learning methods for the prediction of short-term meteorological droughts of sakarya meteorological station in turkey. *Environmental Science and Pollution Research*, 29 (50):75487–75511, 2022. doi: 10.1007/s11356-022-21083-3.

45 Jamshid Piri, Mohammad Abdolahipour, and Behrooz Keshtegar. Advanced machine learning model for prediction of drought indices using hybrid svr-rsm. *Water Resources Management*, 37(2):683–712, 2023. doi: 10.1007/s11269-022-03395-8.

46 A. Belayneh, J. Adamowski, B. Khalil, and J. Quilty. Coupling machine learning methods with wavelet transforms and the bootstrap and boosting ensemble approaches for drought prediction. *Atmospheric Research*, 172-173():37–47, 2016. doi: 10.1016/j.atmosres.2015.12.017.

47 Mohammed Achite, Okan Mert Katipoglu, Serkan Şenocak, Nehal Elshaboury, Ommolbanin Bazrafshan, and Hüseyin Yıldırım Dalkılıç. Modeling of meteorological, agricultural, and hydrological droughts in semi-arid environments with various machine learning and discrete wavelet transform. *Theoretical and Applied Climatology*, 154(1-2):413–451, 2023. doi: 10.1007/s00704-023-04564-4.

48 Guo Chun Wang, Qian Zhang, Shahab S. Band, Majid Dehghani, and Kwok Chau. wing and tho, quan thanh and zhu, senlin and samadianfard, saeed and mosavi, amir (2022). monthly and seasonal hydrological drought forecasting using multiple extreme learning machine models. *Engineering Applications of Computational Fluid Mechanics*, 16(1):1364–1381. doi: 10.1080/19942060.2022.2089732.

49 Shahabbodin Shamshirband, Sajjad Hashemi, Hana Salimi, Saeed Samadianfard, Esmaeil Asadi, Sadra Shadkani, Katayoun Kargar, Amir Mosavi, Narjes Nabipour, and Kwok-Wing Chau. Predicting standardized streamflow index for hydrological drought using machine learning models. *Engineering Applications of Computational Fluid Mechanics*, 14(1):339–350, 2020. doi: 10.1080/19942060.2020.1715844.

50 Mohammed Achite, Muhammad Jehanzaib, Nehal Elshaboury, and Tae-Woong Kim. Evaluation of machine learning techniques for hydrological drought modeling: A case study of the wadi ouahrane basin in algeria. *Water (Switzerland)*, 14(3), 2022. doi: 10.3390/w14030431.

51 Amin Mirboluki, Mojtaba Mehraein, and Ozgur Kisi. Improving accuracy of neuro fuzzy and support vector regression for drought modelling using grey wolf optimization. *Hydrological Sciences Journal*, 67(10):1582–1597, 2022. doi: 10.1080/02626667.2022.2082877.

52 Tribeni C. Sharma and Umed S. Panu. Current state of advances in quantification and modeling of hydrological droughts. *Water (Switzerland)*, 16(5), 2024. doi: 10.3390/w16050729.

53 Mojtaba Shahdad and Behzad Saber. Drought forecasting using new advanced ensemble-based models of reduced error pruning tree. *Acta Geophysica*, 70(2): 697–712, 2022. doi: 10.1007/s11600-022-00738-2.

54 Maryam Mokhtarzad, Farzad Eskandari, and Jamshidi Vanjani. Nima and arabasadi, alireza (2017). drought forecasting by ann, anfis, and svm and comparison of the models. *Environmental Earth Sciences*, 76(21). doi: 10.1007/s12665-017-7064-0.

55 Chaitanya B. Pande, N. L. Kushwaha, Israel R. Orimoloye, Rohitashw Kumar, Hazem Ghassan Abdo, Abebe Debele Tolche, and Ahmed Elbeltagi. Comparative assessment of improved svm method under different kernel functions for predicting multi-scale drought index. *Water Resources Management*, 37(3):1367–1399, 2023. doi: 10.1007/s11269-023-03440-0.

56 Pouya Aghelpour, Yiqing Guan, Hadigheh Bahrami-pichaghchi, Babak Mohammadi, Ozgur Kisi, and Danrong Zhang. Using the modis sensor for snow cover modeling and the assessment of drought effects on snow cover in a mountainous area. *Remote Sensing*, 12(20):1–22, 2020. doi: 10.3390/rs12203437.

57 Ozlem Baydaroglu Yeşilkoy, Kasım Koçak, and Levent Şaylan. Prediction of commonly used drought indices using support vector regression powered by chaotic approach. *Italian Journal of Agrometeorology*, 2020(2):65–76, 2020. doi: 10.13128/ijam-970.

58 Levent Latifoglu and Mehmet Ozger. A novel approach for high-performance estimation of spi data in drought prediction. *Sustainability (Switzerland)*, 15(19), 2023. doi: 10.3390/su151914046.

59 Nazanin Behfar, Elnaz Sharghi, Vahid Nourani, and Martijn J. Booij. Drought index downscaling using ai-based ensemble technique and satellite data. *Theoretical and Applied Climatology*, 155(3):2379–2397, 2024. doi: 10.1007/s00704-023-04822-5.

60 A. Belayneh, J. Adamowski, and B. Khalil. Short-term spi drought forecasting in the awash river basin in ethiopia using wavelet transforms and machine learning methods. *Sustainable Water Resources Management*, 2(1):87–101, 2016. doi: 10.1007/s40899-015-0040-5.

61 K. Sangeetha and K. Mohan Kumar. Flood and drought prediction using the machine learning algorithm support vector regression. *International Journal of Engineering and Advanced Technology*, 9(1):5194–5199, 2019. doi: 10.35940/ijeat.A3001.109119.

62 Mohammed Achite, Somayeh Emami, Muhammad Jehanzaib, Okan Mert Katipoğlu, and Hojjat Emami. An election algorithm combined with support vector regression for estimating hydrological drought. *Modeling Earth Systems and Environment*, 10 (1):1395–1405, 2024. doi: 10.1007/s40808-023-01850-y.

63 Tuba Firdaus, Preeti Gupta, and S. Sangita Mishra. Implementing machine learning models for drought prediction based on metrological drought indices with varying time scales: A case of latur region. *Lecture Notes in Civil Engineering*, 285():183–195, 2023. doi: 10.1007/978-981-19-5077-3-15.

64 Soukayna Mouatadid, Nawin Raj, Ravinesh C. Deo, and Jan F. Adamowski. Input selection and data-driven model performance optimization to predict the standardized precipitation and evaporation index in a drought-prone region. *Atmospheric Research*, 212():130–149, 2018. doi: 10.1016/j.atmosres.2018.05.012.

65 Prabal Das, Sujay Raghavendra Naganna, Paresh Chandra Deka, and Jagalingam Pushparaj. Hybrid wavelet packet machine learning approaches for drought modeling. *Environmental Earth Sciences*, 79(10), 2020. doi: 10.1007/s12665-020-08971-y.

66 A. Belayneh, J. Adamowski, B. Khalil, and B. Ozga-Zielinski. Long-term spi drought forecasting in the awash river basin in ethiopia using wavelet neural networks and wavelet support vector regression models. *Journal of Hydrology*, 508():418–429, 2014. doi: 10.1016/j.jhydrol.2013.10.052.

67 Anteneh Belayneh and Jan Adamowski. Drought forecasting using new machine learning methods. *Journal of Water and Land Development*, 18(9):3–12, 2013. doi: 10.2478/jwld-2013.

68 Anteneh Belayneh, Jan Adamowski, John Quilty, Bahaa Khalil, and Bogdan Ozga-Zielinski. Forecasting drought via bootstrap and machine learning methods. In *Proceedings, Annual Conference - Canadian Society for Civil Engineering*, pages 1456–1465. 2(January, 2013.

69 Amuktamalyada Gorlapalli, Supriya Kallakuri, Pagadala Damodaram Sreekanth, Rahul Patil, Nirmala Bandumula, Gabrijel Ondrasek, Meena Admala, Channappa Gireesh, Madhyavenkatapura Siddaiah Anantha, Brajendra Parmar, Brahamdeo Kumar Yadav, Raman Meenakshi Sundaram, and Santosha Rathod. Characterization and prediction of water stress using time series and artificial intelligence models. *Sustainability (Switzerland)*, 14(11), 2022. doi: 10.3390/su14116690.

70 My Hachem Bekri and El Hmaidi. Abdellah and ousmana, habiba and el faleh, el mati and berrada, mohamed and el aissaoui, kamal and essahlaoui, ali and el ouali, abdelhadi (2021). weather drought index prediction using the support vector regression in the ansegmir watershed, upper moulouya, morocco. *Journal of Water and Land Development*, 50():187–194. doi: 10.24425/jwld.2021.138174.

71 Ali Danandeh Mehr. A novel fuzzy random forest model for meteorological drought classification and prediction in ungauged catchments. *Pure and Applied Geophysics*, 177(12):5993–6006, 2020. doi: 10.1007/s00024-020-02609-7.

72 Geetabai S. Hukkeri, Sujay Raghavendra Naganna, Dayananda Pruthviraja, Nagaraj Bhat, and R. H. Goudar. Drought forecasting: Application of ensemble and advanced machine learning approaches. *IEEE Access*, 1(3):11, 2023. doi: 10.1109/ACCESS.2023.3341587.

73 Siham Acharki, Sudhir Kumar Singh, Edivando Vitor do Couto, Youssef Arjdal, and Ahmed Elbeltagi. Spatio-temporal distribution and prediction of agricultural and meteorological drought in a mediterranean coastal watershed via gis and machine learning. *Physics and Chemistry of the Earth*, 131(), 2023. doi: 10.1016/j.pce.2023.103425.

74 Sedigheh Mohamadi, Saad Sh Sammen, Fatemeh Panahi, Mohammad Ehteram, Ozgur Kisi, Amir Mosavi, Ali Najah Ahmed, Ahmed El-Shafie, and Nadhir Al-Ansari. Zoning map for drought prediction using integrated machine learning models with a nomadic people optimization algorithm. *Natural Hazards*, 104(1):537–579, 2020. doi: 10.1007/s11069-020-04180-9.

75 Omer Coşkun and Hatice Citakoglu. Prediction of the standardized precipitation index based on the long short-term memory and empirical mode decomposition-extreme learning machine models: The case of sakarya, türkiye. *Physics and Chemistry of the Earth*, 131(), 2023. doi: 10.1016/j.pce.2023.103418.

76 Shaoxuan Li, Jiancang Xie, Xue Yang, and Xin Jing. Comparison of hybrid machine learning models to predict short-term meteorological drought in guanzhong region, China. *Water Science and Technology*, 87(11):2756–2775, 2023. doi: 10.2166/wst.2023.162.

77 Shuang Zhu, Xiangang Luo, Si Chen, Zhanya Xu, Hairong Zhang, and Zuxiang Xiao. Improved hidden Markov model incorporated with copula for probabilistic seasonal drought forecasting. *Journal of Hydrologic Engineering*, 25(6), 2020. doi: 10.1061/(ASCE)HE.1943-5584.0001901.

78 Lei Xu, Nengcheng Chen, Xiang Zhang, and Zeqiang Chen. An evaluation of statistical, nmme and hybrid models for drought prediction in China. *Journal of Hydrology*, 566():235–249, 2018. doi: 10.1016/j.jhydrol.2018.09.020.

79 Boksoon Myoung, Jinyoung Rhee, and Changhyun Yoo. Long-lead predictions of warm season droughts in south korea using north atlantic sst. *Journal of Climate*, 33 (11):4659–4677, 2020. doi: 10.1175/JCLI-D-19-0082.1.

80 Narjes Nabipour, Majid Dehghani, Amir Mosavi, and Shahaboddin Shamshirband. Short-term hydrological drought forecasting based on different nature-inspired optimization algorithms hybridized with artificial neural networks. *IEEE Access*, 8(): 15210–15222, 2020. doi: 10.1109/ACCESS.2020.2964584.

81 Kiyoumars Roushangar, Roghayeh Ghasempour, V. S. Ozgur Kirca, and Mehmet Cüneyd Demirel. Hybrid point and interval prediction approaches for drought modeling using ground-based and remote sensing data. *Hydrology Research*, 52(6):1469–1489, 2021. doi: 10.2166/NH.2021.028.

82 K. A. Jariwala and P. G. Agnihotri. Comparative analysis of drought modeling and forecasting using soft computing techniques. *Water Resources Management*, 37(15): 6051–6070, 2023. doi: 10.1007/s11269-023-03642-6.

83 Ali. Danandeh mehr. *and Babak Vaheddoost and Babak Mohammadi ENN-SA: A novel neuro-annealing model for multi-station drought prediction.*, *Computers & Geosciences*, 145(), 104622, 2020. doi: 10.1016/j.cageo.2020.104622.

84 A. Dikshit and B. Pradhan. Interpretable and explainable ai (xai) model for spatial drought prediction. *Science of the Total Environment*, 801, 2021. Art. no. 149594.

85 Masoud Reihanifar and Danandeh Mehr. Ali and tur, rifat and ahmed, abdelkader t. and abualigah, laith and dabrowska, dominika (2023). a new multi-objective genetic

programming model for meteorological drought forecasting. *Water (Switzerland)*, 15 (20). doi: 10.3390/w15203602.

86 G.J. Gomes, R.G. Gomes, and E. do Vargas. A dual search-based epr with self-adaptive offspring creation and compromise programming model selection. *Engineering with Computers*, 38(S3):2155–2173, 2021. doi: 10.1007/s00366-021-01313-x.

87 Moncef Bouaziz, Emna Medhioub, and Elmar Csaplovisc. A machine learning model for drought tracking and forecasting using remote precipitation data and a standardized precipitation index from arid regions. *Journal of Arid Environments*, 189(None), 2021. doi: 10.1016/j.jaridenv.2021.104478.

88 A. Danandeh Mehr. Drought classification using gradient boosting decision tree. *Acta Geophysica*, 69:909–918, 2021.

89 Aki Vehtari, Andrew Gelman, and Jonah Gabry. Practical bayesian model evaluation using leave-one-out cross-validation and waic. *Statistics and Computing*, 27: 1413–1432, 2020.

90 Wes McKinney et al. (2011). pandas: a foundational python library for data analysis and statistics. *Python for high performance and scientific computing*, 14(9):1–9.

91 Pauli Virtanen, Ralf Gommers, and Travis E. Oliphant. and (2020). scipy 1.0: Fundamental algorithms for scientific. *Nature Methods*, 17():261–272.

92 Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, et al. (2011). scikit-learn: Machine learning in python. *The Journal of Machine Learning Research*, 12():2825–2830.

93 Asad Ellahi, Ibrahim M. Almanjahie, Tajammal Hussain, Muhammad Zaffar Hashmi, Shahla Faisal, and Ijaz Hussain. Analysis of agricultural and hydrological drought periods by using non-homogeneous poisson models: Linear intensity function. *Journal of Atmospheric and Solar-Terrestrial Physics*, 198:105190, 2020. doi: 10.1016/j.jastp.2020.105190.

94 R. L. Smith and R. H. Hebbert. A review of recent developments in the calibration and use of radar-based rainfall estimates. *Journal of the American Water Resources Association*, 35(1):5–21, 1999. doi: 10.1111/j.1752-1688.1999.tb03592.x.

95 Z. N. Qaisrani, N. Nuthammachot, K. Techato, Jatoi G. H. Asadullah, B. Mahmood, and R. Ahmed. Drought variability assessment using standardized precipitation index, reconnaissance drought index and precipitation deciles across balochistan, pakistan. *Brazilian Journal of Biology*, 84, 2024. doi: 10.1590/1519-6984.261001. Article e261001.

96 J. R. M. Hosking. On the applicable thresholds for declaring a hydrological drought. In *Journal of the American Water Resources Association*, pages 1133–1144. Wiley Online Library, 35(5, 1999. doi: 10.1111/j.1752-1688.1999.tb03592.x.

97 S. Ahmed and M. Sankaran. *Agricultural drought periods analysis by using nonhomogeneous Poisson models and regionalization of appropriate model parameters.* ResearchGate, 2021. doi: 10.13140/RG.2.2.35363.67863.

98 Douglas Maraun and Martin Widmann. Precipitation downscaling under climate change: Recent developments to bridge the gap between dynamical models and the end user. *Reviews of Geophysics*, 48(3), 2010.

99 Martin A Fischler and Robert C Bolles. Random sample consensus: A paradigm for model fitting with applications to image analysis and automated cartography. *Communications of the ACM*, 24(6):381–395, 1981.

100 Daniel S. Wilks. *Statistical Methods in the Atmospheric Sciences*. Elsevier Academic Press, 3rd edition, 2011.

101 M. A. Alawsi, S. L. Zubaidi, N. S. S. Al-Bdairi, N. Al-Ansari, and K. Hashim. Drought forecasting: A review and assessment of the hybrid techniques and data pre-processing. *Hydrology*, 9(7):115, 2022. doi: 10.3390/hydrology9070115.

102 S. Wang, J. Luo, and S. Yang. Feature selection in machine learning: A new perspective. *Neurocomputing*, 300:70–79, 2020. doi: 10.1016/j.neucom.2018.03.077.

103 Kyle Cranmer et al. Pysr: Fast and flexible symbolic regression in python. *arXiv preprint arXiv:2011.09043*, 2020.

104 John R Koza. *Genetic Programming: On the Programming of Computers by Means of Natural Selection*. MIT press, 1992.

105 Hoshin V Gupta, Harald Kling, Koray K Yilmaz, and Guillermo F Martinez. Decomposition of the mean squared error and nse performance criteria: Implications for improving hydrological modelling. *Journal of Hydrology*, 377(1-2):80–91, 2009.

106 Trent McConaghy. Ffx: Fast, scalable, deterministic symbolic regression technology. In *Genetic Programming Theory and Practice IX*, pages 235–260. Springer, 2011.

107 Christoph Molnar. *Interpretable machine learning. ,* 2020.

108 Sandra Zilker, Sven Weinzierl, Mathias Kraus, Patrick Zschech, and Martin Matzner. A machine learning framework for interpretable predictions in patient pathways: The case of predicting icu admission for patients with symptoms of sepsis. *Health Care Management Science*, pages 1–32, 2024.

109 V. Bento et al. Title of the paper. *Journal Name*, 2021. doi: or.

110 K. Rasouli et al. Title of the paper. *Journal Name*, 2019. doi: or.

111 I. Guyon and A. Elisseeff. An introduction to variable and feature selection. *Journal of Machine Learning Research*, 3:1157–1182, 2003.

112 A. K. Mishra and V. P. Singh. A review of drought concepts. *Journal of Hydrology*, 391(1-2):202–216, 2010. doi: or.

113 K. Deb et al. A fast and elitist multiobjective genetic algorithm: Nsga-ii. *IEEE Transactions on Evolutionary Computation*, 6(2):182–197, 2002.

114 J. Tomasella et al. Title of the paper. *Journal Name*, 2018. doi: or.

APÊNDICE A – Pseudocódigos

Algorithm 1 EPE - Expansão Polinomial Evolutiva

procedure EPE(X, Y, β, N, P)
    Inicialize a população de expressões polinomiais aleatórias Pop
    for cada geração $g = 1$ até N **do**
        for cada indivíduo i em Pop **do**
            Calcule o RMSE da expressão polinomial f_i
            Calcule a penalidade de complexidade
            Calcule o fitness F_i como $F_i = RMSE(f_i) \times (1 + \beta \times pen_i)$
        end for
        Selecione os melhores indivíduos com base em F_i
        Aplique operadores genéticos para gerar nova população Pop
    end for
    retorne a melhor expressão polinomial f_{best} da última geração
end procedure

Algorithm 2 Cálculo do SPI - Índice de Precipitação Padronizada

procedure CalcularSPI(P, m)
    Entrada: Conjunto de dados de precipitação P, período de tempo m
    Saída: Valor do SPI para o período m
    Calcule a média histórica da precipitação X
    Calcule o desvio padrão da precipitação histórica σ
    for cada período $i = 1$ até n **do**
        Calcule a precipitação acumulada X_i para o período m
        Calcule o SPI utilizando a fórmula:
            $SPI_i = \frac{X_i - X}{\sigma}$
    end for
    retorne SPI
end procedure