Rainfall Prediction in Eastern Northeast Brazil Using Machine Learning and Oceanic Predictors

Previsão de Chuvas no Setor Leste do Nordeste Brasileiro com Machine Learning e Preditores Oceânicos

Geraldo Paes[1]*, Pablo Sampaio[1]

Resumo: This study proposes a machine learning (ML) approach to predict rainfall in the eastern sector of Northeast Brazil, a region characterized by significant climatic variability. Using binary classification, models were trained to determine whether four-month period (quadrimester) rainfall would be above or below the historical median. Predictors included oceanic and atmospheric variables (e.g., sea surface temperature, trade winds) identified by previous studies, combined with homogeneous rainfall groups. Data from 1982 to 2023 were divided into quadrimesters (April–July, August–November, December–March) and evaluated using Random Forest (RF), Logistic Regression (LR), K-Nearest Neighbors (KNN), and Support Vector Machine (SVM). Nested cross-validation revealed that RF achieved the highest F1-score (0.671) and recall (0.799) when predicting the rainy quadrimester (April–July), demonstrating strong potential for identifying high-rainfall periods. Despite limited data and high variance, the results underscore ML's viability for rainfall forecasting in the region, offering a baseline for future research with expanded datasets or advanced models.
Keywords: Rainfall prediction — Machine Learning — Northeast Brazil — Climate variables — Homogeneous groups

Resumo: Este estudo propõe uma abordagem de Machine Learning (ML) para prever chuvas no setor leste do Nordeste brasileiro, região marcada por alta variabilidade climática. Utilizando classificação binária, modelos foram treinados para determinar se a precipitação quadrimestral estaria acima ou abaixo da mediana histórica. Foram empregados preditores oceânicos e atmosféricos (e.g., temperatura da superfície do mar, ventos alísios) identificados em estudos anteriores, aliados a grupos homogêneos de precipitação. Dados de 1982 a 2023 foram divididos em quadrimestres (Abril–Julho, Agosto–Novembro, Dezembro–Março) e avaliados com Random Forest (RF), Regressão Logística (LR), K-Nearest Neighbors (KNN) e Support Vector Machine (SVM). Validação cruzada aninhada mostrou que o RF obteve o maior F1-Score (0,671) e recall (0,799) na previsão do quadrimestre chuvoso (Abril–Julho), destacando-se na identificação de períodos de alta precipitação. Apesar da limitação de dados e alta variância, os resultados evidenciam a viabilidade do ML para previsões climáticas na região, servindo como base para pesquisas futuras com dados ampliados ou modelos avançados.
Palavras-Chave: Previsão de chuvas — Machine Learning — Nordeste brasileiro — Variáveis climáticas — Grupos homogêneos

1. Introdução

O Nordeste brasileiro apresenta uma grande diversidade de cenários climáticos ao longo de sua extensão, que refletem as diferentes características de cada parte da região. Por conta disso, ao mesmo tempo em que as regiões do interior podem sofrer com secas prolongadas, que comprometem a agricultura local e o ganha-pão de muitas famílias [1], chuvas intensas, que ocorrem frequentemente na região litorânea, podem causar enchentes com consequências severas para a economia e a segurança da população [2].

Por este motivo, a previsão e o monitoramento das chuvas acabam se tornando muito importantes para ajudar a prevenir os impactos negativos de ambas as situações, além de ajudar no planejamento adequado para a produção agrícola do ano. Com base nessa realidade, este trabalho propõe uma abordagem que utiliza técnicas de Machine Learning (ML), juntamente ao conceito de grupos homogêneos e às variáveis preditoras identificadas por [3], para treinar modelos capazes de prever a precipitação no setor leste do Nordeste brasileiro. Por se tratar de uma investigação inicial, a ideia é aplicar uma estratégia de classificação binária, onde os modelos procuram determinar se um dado período terá precipitação abaixo ou acima da mediana histórica, o que simplifica o problema.

Para isso, baseando-se ainda na definição do período chuvoso da região por [3], os dados coletados foram divididos em três quadrimestres: Abril a Julho (Período Chuvoso), Agosto a Novembro (Período Pós-Chuvoso) e Dezembro a Março (Período Pré-Chuvoso). Com isso, o estudo então avalia o desempenho de quatro modelos de classificação: Random Forest (RF), Logistic Regression (LR), K-Nearest Neighbors (KNN) e Support Vector Machine (SVM) em dois cenários: fazendo a previsão para todos os quadrimestres

e fazendo a previsão apenas para o quadrimestre chuvoso, comparando seus resultados.

O restante deste artigo está organizado da seguinte forma: a seção dois trata dos trabalhos relacionados, a seção três apresenta a metodologia utilizada neste estudo, a seção quatro mostra os resultados em ambos os cenários descritos, a seção cinco discute os resultados e as decisões tomadas ao longo do estudo, e a seção seis apresenta as conclusões e possíveis trabalhos futuros.

2. Trabalhos Relacionados

Estudos e pesquisas a respeito da previsão de chuvas no Nordeste brasileiro não são novidade, e graças a estes, sabe-se que a precipitação nessa região é fortemente influenciada por uma série de fatores meteorológicos, oceânicos e atmosféricos como os ventos alísios, a Temperatura da Superfície do Mar (TSM) e eventos como o El Niño, que em conjunto, resultam em uma grande variabilidade na chuva da região [4].

Uma contribuição particularmente relevante para a área foi feita por [3], cujo trabalho utilizou a chamada Análise de Correlação Canônica (ACC) visando encontrar as melhores variáveis preditoras para as chuvas do setor leste da região. Neste trabalho, os autores definiram o período de abril a julho como o período chuvoso do setor leste, e dividiram as estações pluviométricas do setor em três grupos homogêneos (Grupos 1, 2 e 3) com padrões de precipitação semelhantes devido à geografia. O Grupo 1 é formado por aquelas localizadas nas proximidades ou no próprio litoral, onde ocorrem as maiores chuvas, e é o foco de estudo deste artigo.

[5] também buscaram identificar zonas homogêneas de precipitação, mas dessa vez focando especificamente no estado de Pernambuco. O estudo também denotou a formação de três zonas com características distintas: semiárido, transição e litoral, e verificou uma predominância de anos secos no período analisado (1987 a 2019). Adicionalmente, o trabalho de [6] procurou lidar com a alta variação de chuva da região, usando um modelo empírico para identificar padrões de precipitação ligados às temperaturas de diferentes áreas do oceano, e assim prever a quantidade de chuva em diferentes partes do Nordeste.

Além disso, alguns trabalhos também fazem uso de abordagens baseadas em ML, que podem modelar relações não lineares entre as variáveis ou prever resultados com base em seu treinamento.

Um exemplo é o de [7], que fez uso de Deep Learning ao introduzir uma rede neural convolucional (CNN) projetada para identificar secas repentinas (flash droughts) no Nordeste brasileiro. O modelo foi treinado com dados hidroclimáticos de 2010 a 2022, obtidos do Brazilian Daily Weather Gridded Data (BR-DWGD), gerando mapas probabilísticos de detecção de secas que mostraram uma grande variabilidade espacial, indicando que algumas áreas são mais propensas a secas severas do que outras. Também foram feitas projeções futuras para o período de 2024 a 2050, categorizando diferentes partes da região com base na severidade das secas indicadas nessas projeções.

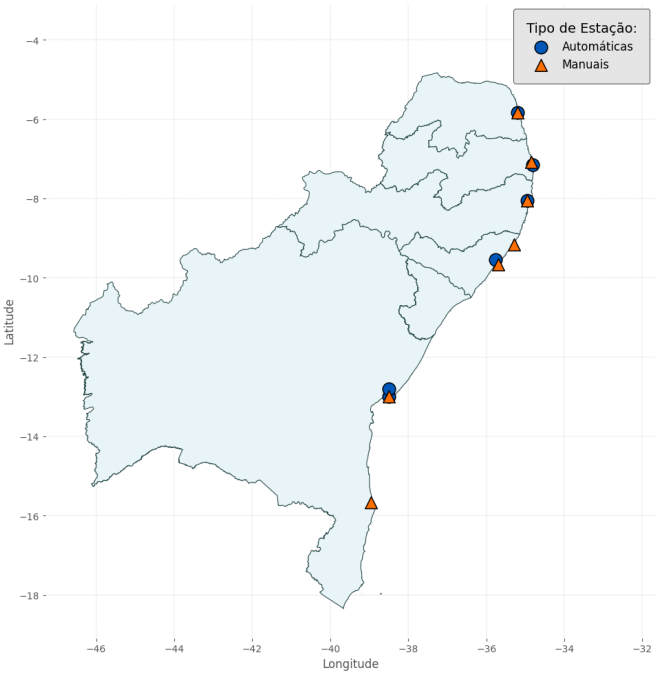

Figura 1. Mapa do Nordeste brasileiro indicando as localizações das estações pluviométricas do grupo homogêneo 1.

Outro estudo pertinente teve foco no Ceará. [8] avaliou vários modelos de ML para prever a precipitação mensal no estado, aplicando teoria do caos e reconstrução do espaço de fase devido às dinâmicas complexas e caóticas das chuvas. O trabalho utilizou dados de 20 estações, dos anos de 1962 a 2006, e testou modelos como Decision Tree, RF, SVM e um modelo ensemble empilhado. Dentre eles, o RF e o ensemble foram os que se saíram melhor, com valores de eficiência Nash-Sutcliffe (NSE) de 0.91 e 0.93, respectivamente.

Ainda na região Nordeste, [9] utilizou modelos GAMLSS (Generalized Additive Models for Location, Scale and Shape) tanto para filtrar os índices com maior eficiência preditiva, quanto para realizar as previsões propriamente ditas, e concentrou seus estudos no estado da Paraíba. E apesar de focar no Sudeste, [10] propôs um modelo baseado em redes LSTM (Long Short-Term Memory) para prever eventos de precipitação extrema na região, algo que pode e já ocorreu no Nordeste, levando a consequências graves, como nos casos das enchentes de 1966 e 1975, ou até mesmo mais recentemente, em 2022 [11].

Embora esses e outros estudos recentes tenham explorado técnicas de ML para a previsão de chuvas, seja no Nordeste ou em outras regiões, nenhum deles integrou os grupos homogêneos e preditores identificados por [3] em suas abordagens. Dessa forma, este trabalho preenche essa lacuna com uma investigação inicial, utilizando os preditores oceânicos e atmosféricos desse estudo como features em

modelos de classificação binária, dados esses que são gratuitos e simples de obter na internet.

3. Metodologia

A proposta deste estudo é realizar uma investigação inicial, com base no trabalho de [3], da previsão de chuva no setor leste do Nordeste brasileiro com a utilização de modelos de ML. Mais especificamente, o foco está nas localidades próximas ou situadas ao longo do litoral, que fazem parte do Grupo 1 segundo a divisão de grupos homogêneos do trabalho (Tabela 1). Optou-se, também, pelo uso de modelos de classificação binária, visando simplificar o processo durante esta primeira abordagem.

Tabela 1. Relação das estações dos grupos homogêneos (recriado de [3]).

Grupo	Estações pluviométricas
Grupo 1	Natal, João Pessoa, Recife, Porto de Pedras, Maceió, Salvador, Canavieiras.
Grupo 2	Ceará Mirim, Areia, Vitória de Santo Antão, Garanhuns, Palmeira dos Índios, Propriá, Aracaju, Itabaianinha, Alagoinha, Cruz das Almas, Guaratinga, Caravelas.
Grupo 3	Campina Grande, Surubim, Arcoverde, Caruaru, Pão de Açúcar, Cipó, Feira de Santana, Itiruçu.

Além disso, considerando a definição do período chuvoso da região como de abril a julho, também por [3], os modelos utilizados não foram treinados para fazer previsões mensais, mas sim dividindo o ano em quadrimestres. Tendo isso em mente, os meses foram divididos da seguinte forma: Abril a Julho (Período Chuvoso), Agosto a Novembro (Período Pós-Chuvoso), e Dezembro a Março (Período Pré-Chuvoso). A partir daí, os modelos usam os dados de um quadrimestre para prever o quadrimestre seguinte, tentando prever se sua precipitação será acima ou abaixo da mediana (por isso, classificação binária). Porém, para evitar vazamento de dados, essa mediana é calculada dinamicamente, processo que é melhor explicado na Seção 3.4.

3.1 Coleta de Dados

Dois tipos de dados foram reunidos para este estudo: os dados de entrada, uma série de variáveis oceânicas e atmosféricas que seriam utilizadas como features dentro dos modelos de classificação; e os dados de saída, se tratando da precipitação total mensal (em milímetros) das várias estações pluviométricas do Grupo 1.

Começando pelos dados de entrada, eles foram obtidos no site da Climate Prediction Center, na aba de Índices Mensais de Atmosfera e Temperatura da Superfície do Mar [12], que disponibiliza gratuitamente arquivos contendo essas informações, com os valores sendo atualizados mensalmente. Esses dados incluem variáveis referentes aos ventos alísios, temperatura da superfície do mar e pressão ao nível do mar em diferentes partes do mundo (descritos na Tabela 2). Os arquivos baixados vieram em formato .txt, cada um contendo tabelas que mostram o valor das variáveis ao longo dos anos e meses. Dados padronizados ou de anomalia presentes nesses arquivos não foram utilizados no estudo.

Os dados de saída, por sua vez, foram coletados do site do Instituto Nacional de Meteorologia (INMET), presentes no Banco de Dados Meteorológicos (BDMEP) [13]. É necessário fazer uma solicitação pelo sistema para receber os dados de estações específicas por e-mail. No caso deste estudo, foram selecionadas todas as estações automáticas e manuais (ou convencionais) das cidades compreendidas no Grupo 1. Aqui, os arquivos vieram no formato .csv, um para cada estação, com um total de 6 automáticas e 7 manuais.

3.2 Pré-processamento

Dado que os arquivos em seu formato original não estavam prontos para serem usados pelos modelos de ML avaliados, foi necessário fazer um tratamento dos mesmos. Esse pré-processamento também foi feito separadamente para os dois tipos de dados, já que os seus formatos e conteúdos eram bastante diferentes.

3.2.1 Dados de Entrada

Os dados de entrada foram divididos em duas categorias, chamadas de Tipo 1 e Tipo 2, com base na sua estrutura.

- Os arquivos do Tipo 1 continham dados de uma única variável, apresentados em várias tabelas. Cada tabela tinha linhas representando anos e colunas representando meses, com cada um desses espaços preenchidos com os valores da variável para os respectivos meses.
- Nos arquivos do Tipo 2, havia apenas uma tabela, onde cada linha representa uma combinação de ano e mês. Assim, cada ano tinha 12 linhas ao todo, e as colunas continham os valores de diferentes variáveis.

Em ambos os casos, foi necessário remover as partes não relevantes do texto, como o cabeçalho que precedia as tabelas e as tabelas adicionais contendo outros tipos de dados, que não seriam usados no estudo. Também foram feitos ajustes nas tabelas, tratando valores vazios e separando os campos de maneira consistente, para evitar que o conteúdo de um campo acabasse vazando para o campo seguinte.

Além disso, os dados do Tipo 1 tiveram que ser transformados para que suas linhas também representassem combinações de ano e mês, assim como nos dados do Tipo 2. Desta forma, eles não teriam problemas de compatibilidade entre si.

3.2.2 Dados de Saída

Nos dados de saída, que tinham uma estrutura mais simples, não foi necessário fazer uma subdivisão dos arquivos. Todos se tratavam de arquivos .csv com informações separadas por ponto e vírgula, e continham um grande número de variáveis,

Tabela 2. Descrição das variáveis oceânicas e atmosféricas utilizadas como *features*.

Variável	Descrição
North Atlantic (5-20°N, 60-30°W)	Temperatura da Superfície do Mar (TSM) no Atlântico Norte.
South Atlantic (0-20°S, 30°W-10°E)	Temperatura da Superfície do Mar (TSM) no Atlântico Sul.
Global Tropics (10°S-10°N, 0-360°)	Temperatura da Superfície do Mar (TSM) nos Trópicos Globais.
Niño 1+2 (0-10°S, 90°W-80°W)	TSM na região do Niño 1+2 (OISST.v2).
Niño 3 (5°N-5°S, 150°W-90°W)	TSM na região do Niño 3 (OISST.v2).
Niño 4 (5°N-5°S, 160°E-150°W)	TSM na região do Niño 4 (OISST.v2).
Niño 3.4 (5°N-5°S, 170-120°W)	TSM na região do Niño 3.4 (OISST.v2).
Tahiti Sea Level Pressure	Pressão ao Nível do Mar (PNM) na estação de Tahiti.
Darwin Sea Level Pressure	Pressão ao Nível do Mar (PNM) na estação de Darwin.
850 MB Trade Wind Index (135E-180W, 5N-5S) – West Pacific	Índice de ventos alísios a 850 hPa no Pacífico Oeste.
850 MB Trade Wind Index (175W-140W, 5N-5S) – Central Pacific	Índice de ventos alísios a 850 hPa no Pacífico Central.
850 MB Trade Wind Index (135W-120W, 5N-5S) – East Pacific	Índice de ventos alísios a 850 hPa no Pacífico Leste.

todas coletadas a partir das medições feitas em diversas estações pluviométricas. No entanto, a maioria delas não era de interesse para este estudo, sendo necessária apenas a precipitação mensal de cada uma.

Assim, para fazer o tratamento desses dados, foi necessário primeiro remover o cabeçalho dos arquivos, que continham várias informações a respeito da estação em questão, como código, coordenadas, se ainda estava em atividade, etc. Em seguida, os dados foram organizados em colunas, e aquelas relacionadas à precipitação foram extraídas. Aqui também foi necessário fazer uma pequena alteração na coluna da "Data de Medição", dividindo-a em duas (uma para o ano e outra para o mês da medição), para que seu formato também fosse compatível com os dados de entrada citados na seção 3.2.1.

3.2.3 Organização dos Dados

Depois de finalizado o tratamento, cada um dos arquivos foi convertido em uma estrutura organizada de dados (um dataframe Pandas) e combinado com os outros do mesmo tipo. Isso gerou dois grandes grupos de dados: um para os de entrada e outro para os de saída.

O conjunto de dados de entrada tinha um total de 12 variáveis oceânicas e atmosféricas, incluindo: índices dos ventos alísios (no Pacífico central, leste e oeste); temperaturas da superfície do mar de diferentes regiões do Pacífico equatorial (Niño 1+2, 3, 4 e 3.4), Atlântico e nos trópicos; e pressão ao nível do mar em Tahiti e Darwin.

Já o conjunto dos dados de saída contava com a precipitação mensal das 13 estações pluviométricas do Grupo 1, com a maioria dos municípios tendo uma ou duas cada. Porém, os períodos dos dados coletados dessas estações variavam bastante, com algumas tendo medições desde 1970, e outras tendo se tornado inativas nesse meio-tempo. Assim, para calcular a precipitação média total de cada mês, foi tirada a média dos valores de todas as estações com dados disponíveis naquele dado período, desconsiderando aquelas que estivessem inativas ou cujos dados ainda não estivessem sendo coletados.

Com os dois conjuntos prontos, o próximo passo foi definir o intervalo dos anos que seriam utilizados no experimento e filtrá-los. Foi selecionado o período de 1982 a 2023, pois todos os arquivos continham informações completas a partir de 1982 (sendo que apenas alguns tinham dados mais antigos), e o ano de 2024 ainda não havia sido concluído na época do estudo, de forma que usá-lo resultaria em dados parciais.

3.3 Transformação de Dados

Em seguida, foi feita a divisão dos dados em quadrimestres, como foi definido no início da Seção 3, com base no valor numérico dos meses de cada linha, e depois agregando todos os dados em um conjunto final. Como o Período Pré-Chuvoso faz uso de um mês do ano anterior (dezembro), o primeiro quadrimestre do conjunto de dados (quadrimestre 1 de 1982) é a única exceção, utilizando apenas os meses de janeiro, fevereiro e março, já que dezembro de 1981 não está presente nos dados. Por conta disso, este quadrimestre foi removido do conjunto final, para evitar o uso de dados incompletos.

No processo de agregar os dados por quadrimestre, foram usados métodos diferentes para as variáveis de entrada e saída. Para as variáveis de entrada, relacionadas às condições oceânicas e atmosféricas, foi calculada a média dos valores ao longo dos quatro meses que compõem cada quadrimestre, oferecendo uma visão geral do comportamento desses preditores naquele período. No caso das variáveis de saída, que se referem à chuva, foi feito o somatório da precipitação total nos quatro meses. Como explicado anteriormente, esse valor era calculado para a região homogênea inteira, usando uma média dos valores de todas as estações disponíveis em um

dado mês. Somando-os dessa forma, foi possível obter uma visão mais clara da quantidade de chuva ocorrida ao longo do quadrimestre inteiro.

Além disso, foram introduzidos em cada quadrimestre um par de indicadores, apontando o ano e o quadrimestre seguinte, de forma a identificar qual período futuro está sendo previsto a partir dos dados atuais. Por exemplo, para o primeiro quadrimestre de um dado ano hipotético, esses campos mostrariam que ele tenta prever o segundo quadrimestre do mesmo ano. Da mesma forma, no caso do terceiro quadrimestre, os indicadores apontariam para o primeiro quadrimestre do ano seguinte.

Ao final desse processo, o resultado é um dataset com um total de 124 instâncias, que é o número de quadrimestres no período escolhido (já contando com a remoção do primeiro quadrimestre de 1982). Isso é de extrema importância, pois significa que os modelos terão uma quantidade bastante pequena de dados para treinar. Cada uma dessas instâncias possui dados organizados como explicado antes nesta seção: a média das variáveis oceânicas e atmosféricas como features de entrada, e a soma das precipitações médias do grupo homogêneo como saída, ambos considerando o período do quadrimestre em questão.

3.4 Treinamento dos Modelos

Para este estudo, foram utilizados quatro modelos de classificação, cada um testando uma série de diferentes hiperparâmetros. Os modelos são: Random Forest (RF), Logistic Regression (LR), K-Nearest Neighbors (KNN) e Support Vector Machine (SVM). A seed utilizada para todos os geradores de números aleatórios foi 0.

O treinamento dos modelos seguiu uma pipeline estruturada, garantindo que todas as transformações e etapas de treinamento fossem aplicadas de maneira consistente. De forma a explicar melhor o processo, esta subseção foi dividida em três partes: transformações em tempo de treinamento, modelos testados e seus hiperparâmetros, e otimização e treinamento.

3.4.1 Transformações em Tempo de Treinamento

Antes da aplicação dos modelos propriamente ditos, os dados precisam passar por algumas transformações importantes para alcançar uma melhora no desempenho e garantir que não haverá vazamento de dados.

- **Divisão das Classes de Saída:** A definição das classes foi feita com base na mediana da precipitação, mas usando apenas os dados do conjunto de treinamento, evitando que os modelos tivessem acesso a informações do conjunto de teste antes da validação. Com isso, os quadrimestres são classificados como acima da mediana (classe positiva) ou abaixo da mediana (classe negativa). Se a mediana global fosse usada em vez disso, os modelos teriam acesso a informações que não deveriam durante o treinamento, implicando em um vazamento de dados. preenchidos com os valores da variável para os respectivos meses.
- **Redução de Dimensionalidade com PCA:** Para avaliar se a redução de dimensionalidade poderia melhorar o desempenho dos modelos, foram testadas duas abordagens. A primeira faz uso de uma transformação que foi chamada de "identidade", que mantém todos os componentes e essencialmente não aplica a redução. A segunda utiliza o PCA, mantendo um número de componentes suficientes para explicar pelo menos 95% da variância dos dados. Dessa forma, os modelos poderiam utilizar a opção que trouxesse os melhores resultados, em vez de obrigatoriamente usar um ou o outro.

3.4.2 Modelos Testados e seus Hiperparâmetros

Cada um dos quatro modelos utilizados no estudo foi testado com um conjunto específico de hiperparâmetros a serem otimizados, conforme detalhado na Tabela 3.

- **Random Forest (RF):** Foram testadas diferentes quantidades de árvores (n_estimators = 10 ou 30) e profundidades máximas (max_depth = 2, 3 ou 4), além de diferentes divisões dos nós (min_samples_split = 2 ou 4). Também foi testado o impacto do balanceamento das classes ao ajustar seus pesos (class_weight = 'balanced' ou None).
- **Logistic Regression (LR):** O hiperparâmetro C foi ajustado entre os valores 0.1, 1 e 10, avaliando diferentes intensidades de regularização. Foram testadas as penalizações L1 e L2, com o solver 'saga' para garantir compatibilidade com ambas as penalizações.
- **K-Nearest Neighbors (KNN):** Foram explorados diferentes números de vizinhos (n_neighbors = 5, 10 ou 20) e pesos de influência (weights = 'uniform' ou 'distance'), para testar se uma ponderação diferente melhoraria a performance do modelo.
- **Support Vector Machine (SVM):** Variou-se o hiperparâmetro C (0.1, 1 ou 10) para controlar a margem de separação. O kernel utilizado foi linear, e aqui também foi avaliado o impacto do balanceamento das classes (class_weight = 'balanced' ou None), como no RF. O parâmetro gamma foi ajustado entre 'scale' e 'auto' para verificar sua influência no ajuste do modelo.

Considerando a pequena quantidade de dados disponíveis, uma combinação mais simples de hiperparâmetros foi escolhida para não aumentar demais a complexidade do treinamento, uma vez que a adição de mais opções, por vezes, causava uma piora no resultado final. Assim, foram estabelecidas, no máximo, duas ou três opções para cada hiperparâmetro.

3.4.3 Otimização e Treinamento

O treinamento foi feito através de uma abordagem chamada validação cruzada aninhada (nested cross-validation), que é muito robusta e permite uma avaliação imparcial dos modelos ao mesmo tempo em que otimiza os seus hiperparâmetros.

- **Validação Cruzada Externa:** O conjunto de dados é dividido em três partes (folds), com duas sendo

Tabela 3. Modelos e hiperparâmetros avaliados.

Modelo	Hiperparâmetros e Valores
RF	pca: Nenhum ou 0.95 n_estimators: 10 ou 30 max_depth: 2, 3 ou 4 min_samples_split: 2 ou 4 class_weight: 'balanced' ou None
LR	pca: Nenhum ou 0.95 C: 0.1,1 ou 10 penalty: 'l1' ou 'l2' solver: 'saga'
KNN	pca: Nenhum ou 0.95 n_neighbors: 5, 10 ou 20 weights: 'uniform' ou 'distance'
SVM	pca: Nenhum ou 0.95 C: 0.1, 1, ou 10 kernel: 'linear' class_weight: 'balanced' ou None gamma: 'scale' ou 'auto'

usadas para treinamento e uma sendo usada para teste, e em seguida iterando entre elas para cobrir todas as possibilidades. De forma a reduzir a variabilidade dos resultados, esse processo foi repetido 30 vezes com diferentes divisões aleatórias dos dados. A estratificação não foi feita com base na classe de saída (que ainda não existia nesse estágio), mas sim nos quadrimestres, assegurando uma distribuição equilibrada.

- **Validação Cruzada Interna e Grid Search:** No loop interno, os dados de treinamento são então subdivididos em outras três partes, seguindo a mesma lógica anterior: duas para treinamento e uma para teste. Em seguida, é aplicado um Grid Search para otimizar os hiperparâmetros de cada modelo dentro dessa subdivisão, buscando a melhor combinação possível.

No fim, cada loop externo resulta em uma instância de modelo treinado que é armazenado em uma lista. Dessa forma, com as trinta repetições, são treinadas no total 90 instâncias de cada modelo, e os resultados finais são as métricas agregadas de todas as execuções armazenadas na lista (médias e desvios padrão). Essa estrutura permite que os modelos sejam avaliados de maneira rigorosa, obtendo uma comparação justa entre eles.

3.5 Avaliação dos Modelos

Para avaliar o desempenho dos modelos, foram escolhidas quatro métricas: F1-Score, Acurácia, Precisão e Recall, com o F1-Score sendo tratado como métrica principal para ranquear os modelos. Com as quatro, é possível fazer uma avaliação mais equilibrada, já que cada uma mede aspectos diferentes da performance do modelo. A Acurácia é a métrica mais direta e simples de entender, representando apenas a taxa de acerto total das predições de um modelo, mas pode ser enganosa a depender do balanceamento das classes. Se um modelo classifica corretamente 6 de 10 previsões, então ele tem uma acurácia de 60%.

As outras métricas são um pouco mais específicas: a precisão representa a quantidade de previsões corretas dentro da classe verdade, ou seja, se o modelo prevê 10 instâncias de "alta precipitação" e acerta 7 delas, sua precisão é de 70%. Por outro lado, o recall representa quantas ocorrências reais dessa classe foram previstas corretamente. Usando o mesmo exemplo, se existem 9 instâncias de quadrimestres com alta precipitação no total e apenas aquelas 7 foram previstas, então seu recall é de aproximadamente 77,8%.

O F1-score, por sua vez, é uma combinação dessas duas métricas, sendo calculado pela média harmônica de precisão e recall. É por esse motivo que ele foi escolhido como a métrica principal na avaliação, já que é uma representação mais balanceada da performance de cada modelo.

Além disso, cada modelo será avaliado em dois cenários diferentes: no Cenário 1, fazendo uma previsão para todos os quadrimestres, de forma a analisar o seu desempenho de maneira geral, e no Cenário 2, tentando prever apenas o quadrimestre chuvoso (Abril a Julho), para investigar se os preditores utilizados têm um impacto maior nesse período.

4. Resultados

Como explicado nas seções anteriores, os modelos avaliados foram RF, LR, KNN e SVM, e as métricas utilizadas foram F1-Score, Acurácia, Precisão e Recall. Os resultados para os dois cenários (previsão de todos os quadrimestres e previsão apenas do quadrimestre chuvoso) serão explicados em mais detalhes nas subseções a seguir.

4.1 Análise do Cenário 1

Analisando a Tabela 4 e a Figura 2, é possível perceber que no cenário onde os modelos tentaram classificar todos os quadrimestres, o desempenho deles, com a exceção do KNN, acabou sendo bastante similar. O SVM e o RF tiveram o exato mesmo F1-Score para 3 pontos flutuantes, 0.605, com desvios padrão ligeiramente diferentes. O LR veio logo em seguida com um score de 0.603, ainda muito próximo dos outros dois, e o KNN ficou em último lugar, com o valor consideravelmente mais baixo de 0.565.

Um ponto que chamou a atenção foi que, embora os F1-Scores tenham sido muito próximos para os três primeiros modelos, o RF foi o único com uma diferença notável entre precisão e recall, com 0.578 e 0.650, respectivamente. Isso mostra que este modelo está conseguindo capturar mais facilmente os exemplos de alta precipitação, ao custo de acabar classificando mais falsos positivos. O SVM e o LR

Tabela 4. Resultados do cenário da previsão de todos os quadrimestres.

Modelo	F1-Score (Média ± DP)	Acurácia (Média ± DP)	Precisão (Média ± DP)	Recall (Média ± DP)
SVM	0.605 ± 0.084	0.612 ± 0.066	0.606 ± 0.095	0.617 ± 0.112
RF	0.605 ± 0.091	0.592 ± 0.073	0.578 ± 0.096	0.650 ± 0.131
LR	0.603 ± 0.078	0.607 ± 0.063	0.603 ± 0.103	0.617 ± 0.100
KNN	0.565 ± 0.096	0.584 ± 0.067	0.589 ± 0.122	0.563 ± 0.127

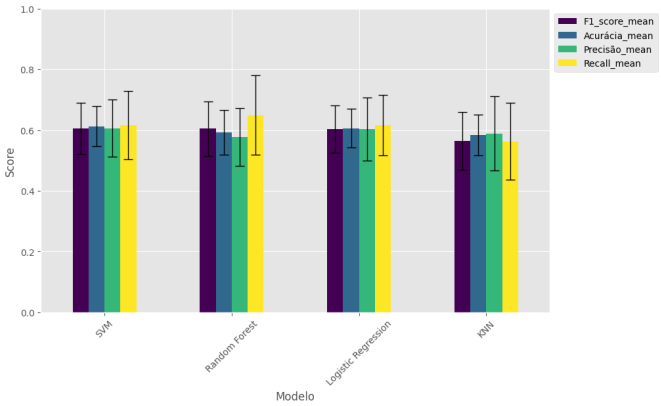

Figura 2. Métricas do cenário da previsão de todos os quadrimestres.

são mais equilibrados nesse sentido, com discrepâncias de apenas cerca de 1 p.p. para sua precisão e recall.

Por outro lado, o classificador KNN teve resultados inferiores em praticamente todas as métricas, mas acabou tendo um resultado inverso ao do RF, ainda que em menor grau: sua precisão é consideravelmente maior que o seu recall, quase se equiparando à precisão dos modelos SVM e LR. O seu desempenho deixou a desejar, com o seu método de utilizar os vizinhos mais próximos para classificar um ponto tendo, possivelmente, sido impactado negativamente pelo conjunto de dados pequeno utilizado, prejudicando sua identificação de padrões.

A otimização de hiperparâmetros realizada durante o treinamento também gerou uma lista com as melhores combinações, baseando-se nos parâmetros mais frequentes dentre as 30 repetições, para cada um dos modelos. Essas informações estão presentes na Tabela 5.

4.2 Análise do Cenário 2

Nesse segundo cenário, é possível notar resultados significativamente diferentes, como mostrado na Tabela 6 e na Figura 3. O melhor modelo, por uma boa margem, foi o RF, alcançando um sólido 0.671 em seu F1-Score, seguido do KNN, que mostrou um desempenho muito superior em relação ao outro cenário, com 0.631. O SVM e o LR tiveram resultados bem parecidos com os do experimento anterior, com 0.613 e 0.607, respectivamente, além de possuírem um maior equilíbrio entre precisão e recall.

Tabela 5. Melhores hiperparâmetros por modelo (cenário 1).

Modelo	Melhores Hiperparâmetros
RF	pca: Nenhum n_estimators: 30 max_depth: 2 min_samples_split: 2 class_weight: 'balanced'
LR	pca: Nenhum C: 10 penalty: 'l1' solver: 'saga'
KNN	pca: Nenhum n_neighbors: 20 weights: 'distance'
SVM	pca: Nenhum C: 10 kernel: 'linear' class_weight: 'balanced' gamma: 'scale'

Mas o que chama mais atenção aqui é o valor do recall do modelo RF, se aproximando dos 80%, se mostrando muito mais alto do que qualquer outra métrica observada até então. Assim como no cenário anterior, isso mostra a alta capacidade do modelo de identificar a classe de alta precipitação. O KNN também se destacou nesse ponto, com um recall de 0.756.

Os hiperparâmetros mais comuns nesse cenário foram os mesmo do cenário anterior para todos os modelos.

5. Discussão

Apesar dos resultados serem relativamente modestos, eles mostram que há bastante potencial no uso de ML, em conjunto com os preditores e grupos homogêneos do estudo de [3], para a previsão de chuvas no setor leste do Nordeste brasileiro. Dada a dificuldade de realizar previsões climáticas somada ao fato de que o estudo se trata de uma investigação

Tabela 6. Resultados do cenário da previsão apenas do quadrimestre chuvoso.

Modelo	F1-Score (Média ± DP)	Acurácia (Média ± DP)	Precisão f(Média ± DP)	Recall (Média ± DP)
RF	0.671 ± 0.132	0.648 ± 0.115	0.607 ± 0.165	0.799 ± 0.159
KNN	0.631 ± 0.171	0.614 ± 0.133	0.573 ± 0.183	0.756 ± 0.217
SVM	0.613 ± 0.134	0.627 ± 0.118	0.619 ± 0.165	0.650 ± 0.185
LR	0.607 ± 0.130	0.607 ± 0.118	0.603 ± 0.183	0.665 ± 0.175

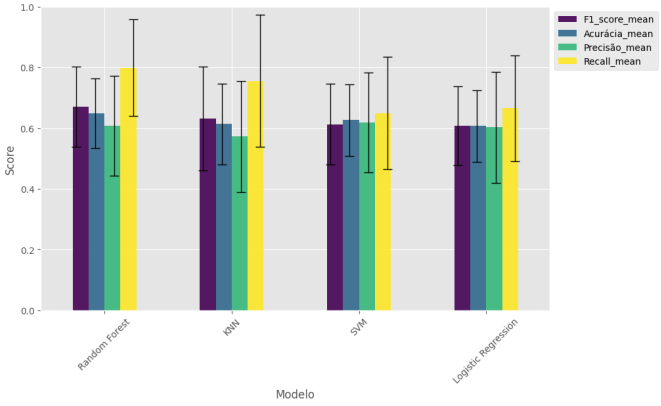

Figura 3. Métricas do cenário da previsão apenas do quadrimestre chuvoso.

inicial, com dados simples e obtidos gratuitamente, o resultado se mostrou positivo, especialmente para o recall, o que superou as expectativas iniciais.

Este estudo foi desenvolvido, desde o início, tendo em mente as suas limitações. Por se tratar de uma investigação inicial do uso dessas variáveis oceânicas e atmosféricas como features num modelo de ML, optou-se por simplificar a abordagem fazendo uso de uma classificação binária. Isso porque a previsão de chuvas é um problema complexo, que depende de muitas variáveis interconectadas, dificultando a modelagem. Mesmo com essas restrições, os resultados mostraram que o uso desses modelos é superior a um classificador aleatório (que teria uma acurácia de aproximadamente 50%), de forma que este estudo pode ser usado como uma referência inicial ou "baseline" para outras pesquisas futuras, que podem incluir novos dados, usar modelos mais robustos ou incorporar novas features.

Um ponto de grande atenção, porém, está no desvio padrão elevado das métricas avaliadas. Devido à quantidade muito limitada de dados, com apenas 124 entradas no conjunto de dados dos quadrimestres, diferentes divisões dos dados (como por exemplo, o uso de diferentes sementes para o gerador aleatório) acabavam gerando uma grande variação nos resultados, e isso acabava sendo representado em desvios padrão elevados. Quando essa variabilidade foi notada pela primeira vez durante os experimentos, foram feitos alguns testes com diferentes valores para a semente aleatória utilizada no código. Foi isso que levou à eventual implementação das múltiplas repetições na validação cruzada, que são responsáveis por mitigar esse problema. Com o uso de suas múltiplas repetições, o resultado obtido não sofreu mais com grandes variações, embora ainda existisse o problema do alto desvio padrão.

Esse efeito é ainda mais exacerbado no segundo cenário, já que a quantidade de dados disponíveis é ainda menor, uma vez que apenas os quadrimestres de Abril a Julho foram utilizados. Para fins de comparação, enquanto o maior desvio padrão no cenário 1 foi de ±0.131, no recall do modelo RF, o cenário 2 chegou a alcançar o valor de ±0.217 no recall do KNN, representando uma variação de mais de 20 p.p. Assim, embora os modelos estejam conseguindo identificar relações entre as variáveis oceânicas e atmosféricas (usadas como features) e a precipitação, o seu desempenho ainda precisa melhorar até que esteja pronto para ser usado em um cenário real.

Mais uma vez, isso traz à tona a questão da quantidade pequena de dados que foram utilizados durante o treinamento e avaliação dos modelos. Uma consideração a se fazer é se uma abordagem diferente no tratamento dos dados de entrada e saída poderia levar a resultados melhores. Afinal, a decisão de dividir os dados em quadrimestres para representar melhor o período chuvoso da região também significa dividir por quatro o número total de instâncias (que originalmente eram meses) disponíveis para os modelos.

6. Conclusão e Trabalhos Futuros

Este trabalho apresentou um estudo inicial voltado para a previsão de chuvas no setor leste do Nordeste brasileiro, demonstrando a viabilidade do uso de modelos de ML em conjunto com os preditores oceânicos e atmosféricos utilizados e nos grupos homogêneos identificados por [3]. Para simplificar o processo, foi implementada uma abordagem de classificação binária, onde os modelos foram treinados para prever se a precipitação de um dado quadrimestre estaria acima ou abaixo da mediana daquele período.

Os dados foram obtidos a partir dos sites do Climate Prediction Center e do INMET, incluindo os índices das variáveis preditoras e os valores de precipitação das estações pluviométricas do Grupo 1. Eles foram processados e organizados em quadrimestres de acordo com a definição do período chuvoso do setor, de abril a julho. A partir daí, os dados já transformados foram utilizados para treinar e avaliar

quatro modelos de ML através de uma validação cruzada aninhada: Random Forest (RF), Logistic Regression (LR), K-Nearest Neighbors (KNN) e Support Vector Machine (SVM), testando uma série de hiperparâmetros para cada um deles a fim de obter a melhor combinação e resultados.

No primeiro cenário, onde foi feita a previsão para todos os quadrimestres, os modelos SVM, LR e RF tiveram um desempenho semelhante, com o F1-Score médio variando entre 0.603 e 0.605, enquanto o do KNN foi apenas de 0.565, tendo uma performance inferior. Já no cenário de previsão voltada especificamente para o quadrimestre chuvoso, os resultados foram mais promissores, com destaque para o RF, que alcançou um F1-Score de 0.671 e um surpreendente recall de 0.799, demonstrando uma alta capacidade de identificar a classe de alta precipitação. Dito isso, o estudo também apresentou limitações. A mais importante foi a quantidade pequena de dados disponíveis para o treinamento dos modelos, ainda mais devido ao tratamento no formato de quadrimestres, que resultou em desvios padrão muito elevados em ambos os cenários.

Dessa forma, este trabalho estabelece uma base inicial que pode ser expandida em pesquisas futuras, seja através do uso de novas fontes de dados ou a aplicação de técnicas que mitiguem os efeitos da quantidade reduzida disponível, seja substituindo a classificação binária e os modelos utilizados por outros métodos de ML, como LSTM e técnicas de precisão de séries temporais.

References

[1] ALVES, J. M. B.; SERVAIN, J.; CAMPOS, J. N. B. Relationship between ocean climatic variability and rain-fed agriculture in northeast brazil. *Climate Research*, v. 38, n. 3, p. 225–236, 2009. Disponível em: ⟨https://doi.org/10.1155/2012/369567⟩.

[2] KOUADIO, Y. K. et al. Heavy rainfall episodes in the eastern northeast brazil linked to large-scale ocean-atmosphere conditions in the tropical atlantic. *Advances in Meteorology*, Wiley Online Library, v. 2012, n. 1, p. 369567, 2012. Disponível em: ⟨https://doi.org/10.1155/2012/369567⟩.

[3] MOURA, G. B. de A. et al. Identificação de preditores para as chuvas do setor leste do nordeste do brasil utilizando análise de correlação canônica. *Revista Brasileira de Geografia Física*, v. 13, n. 04, p. 1463–1482, 2020. Disponível em: ⟨https://www.academia.edu/download/69628405/36005.pdf⟩.

[4] HOUNSOU-GBO, G. A. et al. Tropical atlantic contributions to strong rainfall variability along the northeast brazilian coast. *Advances in meteorology*, Wiley Online Library, v. 2015, n. 1, p. 902084, 2015. Disponível em: ⟨https://onlinelibrary.wiley.com/doi/pdf/10.1155/2015/902084⟩.

[5] BEZERRA, A. C. et al. Annual rainfall in pernambuco, brazil: Regionalities, regimes, and time trends. *Revista Brasileira de Meteorologia*, SciELO Brasil, v. 36, n. 3, p. 403–414, 2021. Disponível em: ⟨https://www.scielo.br/j/rbmet/a/fNWxmtcdDrYzfswrbXVGDdd/?format=pdf&lang=en⟩.

[6] JÚNIOR, R. L. da R. et al. An empirical seasonal rainfall forecasting model for the northeast region of brazil. *Water*, Multidisciplinary Digital Publishing Institute, v. 13, n. 12, p. 1613, 2021. Disponível em: ⟨https://www.mdpi.com/2073-4441/13/12/1613⟩.

[7] BARBOSA, H. A.; BURITI, C. O.; KUMAR, T. L. Deep learning for flash drought detection: A case study in northeastern brazil. *Atmosphere*, MDPI, v. 15, n. 7, p. 761, 2024. Disponível em: ⟨https://www.mdpi.com/2073-4433/15/7/761/pdf⟩.

[8] ROLIM, L. Z. R.; FILHO, F. d. A. de S. Machine learning strategies for multiannual rainfall prediction and drought early warning: insights from ceará, brazil. *Natural Hazards*, Springer, p. 1–35, 2024.

[9] DANTAS, L. et al. *Rainfall Prediction in the State of Paraíba, Northeastern Brazil Using Generalized Additive Models. Water, 12, Article 2478*. 2020. Disponível em: ⟨https://www.mdpi.com/2073-4441/12/9/2478⟩.

[10] ARAÚJO, A. de S.; SILVA, A. R.; ZÁRATE, L. E. Extreme precipitation prediction based on neural network model–a case study for southeastern brazil. *Journal of Hydrology*, Elsevier, v. 606, p. 127454, 2022.

[11] FOLHA DE PERNAMBUCO. *Maior tragédia do século em Pernambuco, mortes pelas chuvas de 2022 superam total da cheia de 1975*. Acessado em: 18/03/2025. Disponível em: ⟨https://www.folhape.com.br/noticias/maior-tragedia-do-seculo-em-pernambuco-mortes-pelas-chuvas-de-2022/228963/⟩.

[12] CLIMATE PREDICTION CENTER. *Monthly Atmospheric and SST Indices*. Acessado em: 18/03/2025. Disponível em: ⟨https://www.cpc.ncep.noaa.gov/data/indices/⟩.

[13] INSTITUTO NACIONAL DE METEOROLOGIA. *Banco de Dados Meteorológicos do INMET*. Acessado em: 18/03/2025. Disponível em: ⟨https://bdmep.inmet.gov.br/⟩.