Algoritmos de Aprendizado de Máquina Aplicados na Avaliação do Volume de Chuvas na Cidade do Recife

Adriano Severiano de Albuquerque
Programa de Pós-Graduação em Informática Aplicada (PPGIA)
Universidade Federal Rural de Pernambuco (UFRPE)
Recife, Brasil
adriano.severiano@ufrpe.br

Maria Beatriz Arruda Vieira de Lima
Departamento de Computação (DC)
Universidade Federal Rural de Pernambuco (UFRPE)
Recife, Brasil
mariabeatriz.lima@ufrpe.br

Renata Freire de Paiva Neves
Instituto Federal de Pernambuco (IFPE)
Recife, Brasil
renatafreire@recife.ifpe.edu.br

Danilo Ricardo Barbosa de Araujo
Departamento de Computação (DC)
Universidade Federal Rural de Pernambuco (UFRPE)
Recife, Brasil
danilo.araujo@ufrpe.br

***Resumo*—A crise climática tem cada vez mais relevância devido aos seus eventos extremos e constantes, com diversos impactos sociais e econômicos. A previsibilidade destes eventos, por sua vez, é cada vez mais desafiante. Os avanços em Inteligência Artificial (*IA*) possibilitam a extração e classificação de informações capazes de serem usadas na modelagem de dados meteorológicos, e com isto, auxiliam na mitigação dos impactos destes eventos extremos. Este artigo apresenta um comparativo entre modelos de aprendizado de máquina para a classificação do volume de chuvas na cidade do Recife, por meio da análise de dados semanais de precipitação, com objetivo futuro de oferecer maior assertividade na emissão de alertas de eventos extremos. São considerados os seguintes algoritmos: *k-nearest neighbors*, *logistic regression*, *support vector machine*, *decision tree* e *random forest*. Foram utilizados dados históricos no intervalo de 17 anos, entre 2005 e 2021, extraídos da estação automática (A301) da base do Instituto Nacional de Meteorologia (INMET), localizada em Recife. De acordo com os resultados obtidos, o modelo baseado em regressão logística se destacou, com acurácia de 94,12%. Além disso, foram consideradas também as métricas *recall score* e *receiver operating characteristic score*.**

***Palavras-chave*–Inteligência Artificial, Aprendizado de Máquina, Algoritmos de Classificação, Aprendizado Supervisionado, Alertas de Chuvas Intensas, Mitigação de Eventos Extremos.**

I. Introdução

Devido às mudanças climáticas, a frequência, intensidade e duração de eventos extremos têm aumentado nos últimos anos, incluindo inundações, secas, tempestades e temperaturas extremas. Esses eventos críticos frequentemente resultam em custos financeiros, perda de propriedades e até mesmo vidas, especialmente em países pobres e em desenvolvimento, nos quais não há ou há má gestão de desastres por dificuldades sociais e financeiras ou por falta de sinergia entre as ações de mitigação realizadas. Nos últimos 30 anos, uma análise da ocorrência de chuvas de alta intensidade e secas históricas mostram que eles se tornaram mais frequentes [1]. A cidade do Recife, localizada em Pernambuco, Brasil, pode ser considerada uma região crítica quando se trata de ameaças relativas aos efeitos do clima. O relatório do Painel Intergovernamental das Mudanças Climáticas (IPCC) de 2022 [2] aponta que Recife é a 16ª cidade mais ameaçada do mundo em termos de eventos críticos relacionados com mudanças climáticas. Quando se tratam de acúmulos de chuvas, o relatório Análises de Riscos e Vulnerabilidades Climáticas e Estratégia de Adaptação do Município do Recife [3] aponta que indicadores referentes a maior precipitação em um dia e acumulado de cinco dias mostram tendências de aumento e recorrência de eventos de inundação e deslizamentos, uma vez que as chuvas serão mais intensas e concentradas em um curto período.

Conforme [4], modelos hidrológicos são classificados, dentre outras formas, de acordo com os tipos de variáveis utilizadas na modelagem, os tipos de relações entre essas variáveis, a forma de representação dos dados, a existência ou não de relações espaciais e a existência de dependência temporal. Os avanços científicos e tecnológicos atuais estão viabilizando modelos com baixo custo de implementação e baixa exigência de recursos e capacidade computacional. O trabalho [5] aponta que o tempo computacional de um modelo empírico, por

exemplo, é menor em relação aos modelos convencionais, pois a sua resolução é mais baixa, o que resulta em uma menor quantidade de dados para seu desenvolvimento. O uso de técnicas de Inteligência Artificial em hidrometereologia pode ser considerado uma medida disruptiva para os avanços na área. Segundo [6], um modelo de previsão eficiente baseado nos algoritmos de aprendizado de máquina é capaz de lidar com informações imprecisas e ruidosas sem efeito negativo perceptível na qualidade da resposta, conseguindo identificar padrões existentes entre as amostras desconhecidas.

Este trabalho investiga qual o melhor algoritmo de classificação para identificação de volumes de chuvas críticos, aplicados em dados da cidade do Recife. De acordo com o levantamento bibliográfico realizado, embora existam alguns trabalhos em temática correlata, não há estudos conclusivos sobre as melhores técnicas de classificação para o cenário considerado no estudo. Neste trabalho foram considerados cinco algoritmos de aprendizado de máquina, a saber: *k-nearest neighbors (KNN)*, *logistic regression (LR)*, *support vector machine (SVC)*, *decision tree (DT)* e *random forest (RF)*. As seguintes métricas foram usadas para comparação dos algoritmos: *accuracy score*, *recall score* e *receiver operating characteristic score (ROC Score)*. Os dados usados foram extraídos do Instituto Nacional de Meteorologia (INMET) [7], considerando dados de precipitação semanal nos anos de 2005 a 2021. Os resultados obtidos neste trabalho poderão ser usados futuramente para viabilizar sistemas mais precisos para emissão de alertas de eventos críticos decorrentes de precipitações intensas no cenário considerado no estudo.

Este artigo está organizado da seguinte forma: a Seção II aborda os principais trabalhos relacionados ao estudo em questão, a Seção III explica os conceitos básicos necessários ao entendimento deste trabalho. A Seção IV apresenta os procedimentos metodológicos adotados no estudo, incluindo a explicação dos experimentos realizados. A Seção V apresenta e discute os resultados. A Seção VI apresenta a conclusão do estudo e sugestões para trabalhos futuros.

II. Trabalhos Relacionados

O fenômeno de precipitação tem sido explorado nos últimos anos por técnicas de Inteligência Artificial em diversos estudos [8], seja para classificar ou prever eventos meteorológicos. É recorrente a tentativa de inferir novas informações a partir de amostras de dados históricos e métodos estatísticos são levados em consideração no processo de análise dos resultados obtidos pelos diferentes modelos [9]. Nos parágrafos seguintes serão apresentados brevemente os principais estudos que estão alinhados com a temática do presente trabalho.

O artigo [10] explorou 5 (cinco) algoritmos para desenvolver um sistema de classificação de inundações para o estado de Kerala (Índia). O estudo fez uso do conceito de Inteligência Artificial Explicável (*XAI*) [11]. Os autores avaliaram a precisão das técnicas, bem como a validade das descobertas com base nos dados históricos mensais de precipitação. Os dados foram coletados para um período de 116 anos e a Regressão Logística ofereceu o modelo com os melhores resultados. Percebe-se ainda no artigo que os autores fazem uma análise sobre a perspectiva de dados de longo prazo, deixando uma lacuna para cenários de curto prazo. Como proposta futura, eles se propõem a considerar modelos de aprendizagem profunda (*Deep Learning*) [12], no contexto de classificação de séries temporais [13] e interação homem-máquina [14], visando uma estrutura que possibilite aos usuários encontrar uma solução iterativa que auxilie no processo de previsão de inundação.

O trabalho [15] realizou um comparativo entre algoritmos de *Machine Learning* aplicados na ocorrência de chuvas na cidade de Santa Maria – RS, sendo eles: *support vector machine (SVM)*, *k-nearest neighbors (KNN)*, *classification and regression trees (CART)* e *random forest (RF)*. Foi utilizada uma base de dados do INMET contendo 8.694 observações em um período de 1 ano (setembro de 2018 a setembro de 2019). O objetivo do artigo foi tratar de aprendizado supervisionado referente a variáveis de classificação. Buscou-se a classificação de um período de chuva relacionada às condições climáticas no horário anterior. Dos modelos aplicados, o que obteve o melhor desempenho foi o *random forest (RF)* obtendo 94,84% na validação e 94,13% na base de teste. Contudo, o modelo escolhido foi a *classification and regression trees (CART)* por exigir baixo custo computacional e ser de fácil implementação, tendo também proporcionado uma maior precisão referente aos dados de teste 94,74%. É importante mencionar que a escolha do modelo *CART* não deve ser considerado como um algoritmo definitivo para a previsão de chuvas, tendo em vista que sua escolha foi diretamente relacionada ao banco de dados considerado. O clima da cidade de Santa Maria – RS é subtropical, enquanto que o clima da cidade do Recife tropical-úmido, sendo este último menos previsível e mais desafiante. Ou seja, deve-se levar em consideração que cada região possui condições climáticas diferentes e com outros desafios, que exigem estudos específicos.

No artigo [16] os autores se propõem a investigar diferentes técnicas de aprendizado de máquina visando prever um volume anual de chuva na cidade de Manaus. Foram considerados registros de 65 anos de precipitação mensal e índices *Niño*, segundo quatro abordagens: árvores de decisão, florestas aleatórias, redes neurais e vizinhos mais próximos. Como atributos, foram consideradas medidas de precipitação mensal obtidas de estações meteorológicas automáticas do INMET localizadas em Manaus e dados de temperatura na superfície do mar em diversos pontos do Oceano Pacífico, caracterizando índices *Niño* e por estarem fortemente associados à variabilidade da chuva na bacia amazônica. Por fim, também foi utilizado a anomalia de precipitação dos anos anteriores. Os resultados obtidos mostraram que, embora todos os modelos contemplados pudessem endereçar o problema de previsão, as melhores métricas de desempenho foram percebidas nas redes neurais artificiais obtendo-se um *F-score* de 70%. Ainda que diferentes técnicas tenham sido consideradas, os resultados obtidos por elas não superaram as verificadas nas redes neurais artificiais. Tal constatação reforçou a importância deste modelo na realização de tarefa não-trivial.

O estudo [19] utilizou o algoritmo *k-nearest neighbors*

(KNN) para viabilizar a previsão de riscos de alagamentos e inundações na região metropolitana de São Paulo. Para isto, foram coletados dados de algumas fontes como: radares meteorológicos, satélite, dados históricos, além de mapa de suscetibilidade física aos alagamentos e inundações da região de estudo proposta. Como resultado, o trabalho apresentou ótima perspectiva na identificação dos riscos de eventos hidrometeorológicos severos. Contudo, os autores chegaram a conclusão que o algoritmo *KNN* é computacionalmente dispendioso quando aplicado a grande base de dados, tendo em vista que a cada nova amostra o algoritmo calcula a distância entre ela e todos os elementos pertencentes ao conjunto de treinamento.

Ainda que vários trabalhos envolvendo a temática em questão possam ser citados, determinar o melhor modelo de classificação ou previsão de inundações para uma região específica e em cenários de curto prazo é uma tarefa complexa e ainda em aberto. Este trabalho analisa qual o melhor algoritmo de classificação para identificação de volumes de chuvas críticos, aplicados em dados da cidade do Recife, que apresenta clima tropical-úmido, que é um cenário notadamente desafiante. Além disso, a cidade do Recife apresenta características geográficas e de urbanização peculiares como estar situada no nível do mar, não possuir rede eficiente de esgoto (a rede de águas pluviais é inclusive usada para esgoto) e nem manejo das águas da chuva. Essas características dificultam ainda mais a construção de um sistema de alertas eficiente para inundações e justificam estudos que priorizem este tipo de cenário.

III. Fundamentação Teórica

Nesta seção são explorados alguns conceitos básicos a respeito das técnicas de aprendizagem de máquina consideradas neste estudo e as métricas utilizadas para avaliação.

A. Técnicas de Aprendizagem de Máquina

Modelos de Aprendizado de Máquina (*ML)* visam encontrar padrões para tomada de decisão a partir de um conjunto de dados. Durante a fase de treinamento, o algoritmo é otimizado de modo a encontrar estruturas e regras específicas a depender da tarefa para o qual foi designado. Em um contexto geral, algoritmos de (*ML)* estão contidos dentro do universo da Inteligência Artificial (*IA*), e são métodos que se utilizam de expressões matemáticas e que são derivados do campo da Estatística, Cálculo e Álgebra Linear. As técnicas podem ser classificadas em aprendizado supervisionado, aprendizado não supervisionado e aprendizado por reforço. O aprendizado supervisionado consiste em usar um rótulo para os dados previamente estabelecidos, ou seja, em termos de classificação, já se tem esses dados classificados e rotulados, de modo que o algoritmo vai buscar, a partir deste momento, aprender qual o significado e a relação desses dados. Nos parágrafos seguintes são brevemente explicados os algoritmos que foram usados neste estudo.

K-nearest neighbors (KNN): a regra dos K-vizinhos mais próximos é um algoritmo utilizado como ferramenta de classificação de padrões. De forma mais simples, é uma técnica que busca classificar cada amostra de um conjunto de dados avaliando sua distância em relação aos seus vizinhos mais próximos. A partir disso, caso os vizinhos mais próximos forem predominantemente de uma classe, a amostra em questão será classificada nessa categoria. Segundo [19], apenas dois fatores precisam ser definidos para aplicação desta metodologia: o parâmetro k, que indica o número de vizinhos que serão considerados; e a métrica para o cálculo da similaridade entre um ponto e todos os outros pertencentes ao conjunto de treinamento.

Logistic Regression (LR): é um algoritmo de aprendizado de máquina bastante utilizado para a classificação de dados, e é constantemente utilizado para análise preditiva. De acordo com [20], o modelo de regressão logística adota variáveis dependentes categorizadas de forma binária para identificação do evento que interessa. Com isso, é uma técnica bastante utilizada quando se busca uma saída com 0 ou 1, verdadeiro ou falso, sim ou não, dentre outras categorizações binárias.

Support Vector Machine (SVM): é um algoritmo que analisa dados e reconhece padrões e pode ser utilizado para problemas de classificação ou regressão. Essa técnica encontra um hiperplano, o qual separa os dados em duas classes, buscando maximizar a distância entre os pontos mais próximos a cada uma das classes. Dessa forma, a distância entre o hiperplano e o primeiro ponto de cada uma das classes é conhecido como margem, diante disso a *SVM* primeiro classifica as classes e logo após, a partir dessa classificação define a distância entre as margens. De acordo com [21], tendo como entrada um conjunto de dados, o *SVM* padrão, prediz para qual, das duas possíveis classes, cada ocorrência pertence.

Decision Tree (DT): é um algoritmo de aprendizado de máquina supervisionado que pode ser utilizado para problemas do tipo classificação e regressão. Utiliza uma abordagem de divisão de conquista para classificar eventos usando uma representação baseada em árvores. Uma árvore de decisão é um modelo representado graficamente por nós e ramos. O nó raiz é o primeiro nó da árvore e fica no topo da estrutura. Cada nó contém um teste sobre um ou mais atributos (parâmetros) e os resultados deste teste formam os ramos das árvores [22]. Cada nó folha, nas extremidades da árvore, representa um valor de predição para o atributo meta [23].

Random Forest (RF): é um algoritmo de aprendizado de máquina bastante flexível, podendo ser utilizado para problemas de classificação ou regressão. Esse algoritmo, de forma mais simples, cria diversas árvores de decisão, fazendo uma combinação entre elas para obter uma predição com maior precisão. Esse método é conhecido como o método *bagging*, no qual consiste em realizar uma combinação das árvores gerando uma boa generalização. Segundo [24], uma das diferenças entre a floresta aleatória e a árvore de decisão é que a floresta aleatória é facilmente generalizada e impede o *overfitting*, pois cria subconjuntos aleatórios dos recursos, construindo árvores menores.

B. Métricas avaliadas

Accuracy Score: em geral, a acurácia pode ser considerada como a quantidade de acertos de um modelo, ou seja, o número de previsões corretas (NPC), dividido pelo total de amostras ou número total de previsões (NTP) (Eq. 1). Com ela se quer saber o quão certo (preciso) o modelo está. Na classificação multirótulo, a pontuação de precisão é calculada a partir do subconjunto, ou seja, o conjunto de rótulos previstos para uma amostra e que deve corresponder exatamente ao conjunto correspondente de rótulos.

$$Accuracy = \frac{NPC}{NTP}. \quad (1)$$

Recall Score: seu objetivo é medir a quantidade de vezes que o modelo acerta em relação ao total de vezes que ele deveria ter acertado. Também pode ser conhecido como taxa de detecção e é representado como o número de previsões positivas corretas (NPPC) dividido pelos exemplos positivos (EP) (Eq. 2).

$$Recall = \frac{NPPC}{EP}. \quad (2)$$

Receiver Operating Characteristic (ROC Score): informa qual é a probabilidade de que uma instância positiva escolhida aleatoriamente tenha uma classificação mais alta do que uma instância negativa escolhida aleatoriamente. Geralmente é utilizado quando se está preocupado em procurar as previsões de classificação e não necessariamente com a saída das probabilidades bem calibradas. Segundo [31], o *ROC* possui dois parâmetros: a taxa de verdadeiro positivo (Eq. 3), que é dado pelos verdadeiros positivos (VP) dividido pelos verdadeiros positivos (VP), mais os falsos negativos (FN); e a taxa de falso positivo (Eq. 4), que é dado pelos falsos positivos (FP) divididos pelos falsos positivos (FP), mais os verdadeiros negativos (VN).

$$(i)\ Taxa\ de\ Verdadeiro\ Positivo = \frac{VP}{VP + FN}. \quad (3)$$

$$(ii)\ Taxa\ de\ Falso\ Positivo = \frac{FP}{FP + VN}. \quad (4)$$

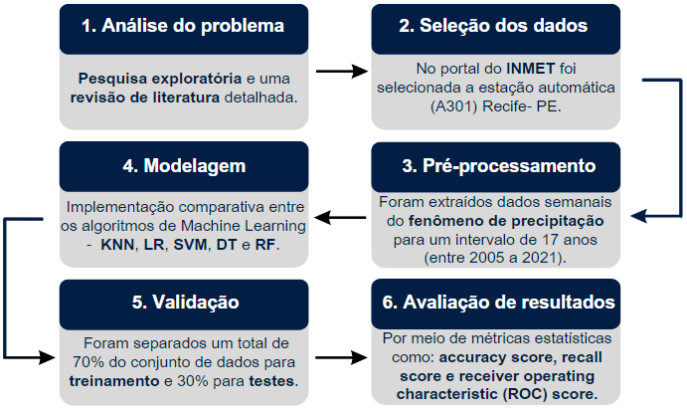
IV. Metodologia

Este estudo foi conduzido considerando experimentos computacionais sobre uma base de dados publicamente disponível, com o objetivo de determinar o melhor modelo de classificação para volumes de chuva na cidade do Recife. A Figura 1 ilustra, de forma geral, os principais passos adotados no estudo, a saber: (i) análise do problema; (ii) seleção dos dados; (iii) pré-processamento; (iv) modelagem; (v) validação; e (vi) avaliação dos resultados.

Cada um destes passos é detalhado a seguir:

(i) *Análise do problema*: mediante uma pesquisa exploratória, buscou-se na literatura estudos relacionados a eventos críticos de precipitação, com foco nos que faziam uso principalmente de técnicas de *IA* para classificação e posterior mitigação de riscos em curto prazo. Foram considerados

Figura 1. Fluxograma da metodologia aplicada.



trabalhos no intervalo de 2012 a 2023, considerando aderência com a temática e o cenário deste estudo.

(ii) *Seleção dos dados*: foram consideradas inicialmente diversas fontes de dados, incluindo a base do Departamento de Controle do Espaço Áereo (DECEA), base do INMET e dados de estações metereológicas próprias adquiridas para o projeto de pesquisa. Considerando aspectos de delimitação de escopo e reprodutibilidade do estudo, a análise comparativa deste artigo focou na estação automática de Recife-PE (A301), cuja extração dos dados se deu por meio do portal do INMET. Foram coletadas informações para um intervalo de 17 anos (2005 a 2021). Trabalhos futuros pretendem utilizar as outras fontes de dados e cruzamento entre estas fontes.

(iii) *Pré-processamento*: da base em questão, foram extraídos como variáveis, dados semanais do fenômeno de precipitação dentro do intervalo de anos proposto. Nesta fase houve uma preparação dos dados partindo da organização dos valores acumulados de chuva para cada semana dos respectivos meses de cada ano. Foi definido um valor a partir da mediana de todos os valores semanais e o máximo valor de cada semana, obtendo assim um valor comparativo para o limiar de acúmulo do volume de chuvas.

(iv) *Modelagem*: os modelos analisados são obtidos a partir de cinco algoritmos de *ML*, a saber: *KNN*, *LR*, *SVM*, *DT* e *RF*. Todos os parâmetros utilizados nos algoritmos ora citados foram obtidos empiricamente. A biblioteca *Scikit-learn*, utilizada para desenvolver o modelo de ML proposto, utiliza a normalização (*MinMaxScaler*), também conhecida como escalonamento mínimo-máximo, em que os valores em uma coluna são deslocados de modo que fiquem limitados entre um intervalo fixo de 0 e 1. Para o *KNN*, a classe mais próxima será identificada usando medidas de distância euclidiana, onde o valor de K é 1 (K = 1), neste caso, a nova classe alvo do ponto de dados será atribuída ao primeiro vizinho mais próximo. Depois de treinados, todos os demais algoritmos utilizam o método predict(*X_test*) para obter resultados de previsão, onde se comparam com os valores verdadeiros de y em (*y_test*). Ao dimensionar o conjunto de dados, utiliza-se da validação cruzada por meio das funções auxiliares (*cross_val_score e cross_val_predict*) onde os dados são divididos de acordo

com o parâmetro cv. No caso de uma ou mais classes estarem ausentes em uma parte do treinamento, uma pontuação padrão precisa ser atribuída a todas as instâncias dessa classe, para o método (*predict_proba*) esse valor é 0. Por fim, é realizada uma comparação de desempenho dos resultados de classificação relacionados ao acúmulo do volume de chuvas, considerando as métricas descritas em (vi).

(v) *Validação*: para esta etapa, foi separado do conjunto total da base coletada, em uma proporção de 70% para dados de treinamento e de 30% para dados de teste.

(vi) *Avaliação de resultados*: o comparativo de resultados entre as técnicas aplicadas se deu por meio da análise das métricas estatísticas *accuracy score, recall score* e *receiver operating characteristic (ROC) score*.

A. Cenário Considerado no Estudo

Recife, como diversas outras cidades do Brasil, tem testemunhado ao longo dos últimos anos alterações extremas no clima. Como parte das preocupantes consequências, destaca-se o avanço do nível do mar, tornando-a uma das cidades mais vulneráveis. Segundo noticiário da Rádio Agência Nacional [17], somente em fevereiro de 2023, o volume de chuva em Recife chegou a 130% do previsto, o que foi suficiente para causar alagamentos, deslizamentos e mortes. Além disso, conforme a reportagem da Folha de Pernambuco [18], em 25 de maio do ano de 2022, foi registrado em apenas 24h, um acumulado de chuva de 197,70 mm, fazendo deste, o terceiro maior acúmulo registrado na cidade nos últimos 50 anos.

B. Detalhamento dos Dados Considerados

Para análise do fenômeno mencionado neste trabalho, a base de referência utilizada no modelo foi coletada da estação automática (A301) para a Cidade do Recife em um período de 17 anos, entre 2005 e 2021 por meio do portal do Instituto Nacional de Meteorologia (INMET) [7].

Figura 2. Dados semanais da estação (A301) Recife-PE - INMET.

	LOCALIDADE	ANO	SEMANA	JAN	FEV	MAR	ABR	MAI	JUN	JUL	AGO	SET	OUT	NOV	DEZ	MEDIANA	MAXIMO	Valor entre MED_MAX	ACUMULO_CHUVA
0	RECIFE	2005	1	16.8	12.2	0.0	35.0	31.0	244.8	31.4	56.6	14.2	44.4	7.0	1.8	14.8	244.8	129.8	SIM
1	RECIFE	2005	2	0.2	36.6	1.0	21.8	107.8	106.2	51.0	18.0	24.4	8.2	1.0	141.6	14.8	141.6	78.2	NAO
2	RECIFE	2005	3	0.8	20.0	12.0	52.6	149.0	242.2	13.2	107.2	6.0	5.8	1.6	6.4	14.8	242.2	128.5	SIM
3	RECIFE	2005	4	0.0	17.6	32.0	14.6	190.6	94.4	50.8	110.2	0.6	0.0	1.0	2.0	14.8	190.6	102.7	SIM
4	RECIFE	2005	5	0.0	0.0	36.2	0.0	47.4	44.2	21.4	14.8	1.0	0.0	0.4	13.6	14.8	47.4	31.1	NAO
80	RECIFE	2021	1	1.2	1.2	73.2	26.6	127.4	30.0	57.4	99.8	10.0	10.4	0.6	0.0	14.8	127.4	71.1	NAO
81	RECIFE	2021	2	2.2	29.4	2.4	246.4	244.0	27.4	20.0	121.4	4.6	3.4	7.6	0.0	14.8	246.4	130.6	SIM
82	RECIFE	2021	3	14.4	61.2	42.8	100.2	11.0	73.8	49.0	54.6	13.4	9.8	0.0	0.0	14.8	100.2	57.5	NAO
83	RECIFE	2021	4	23.2	28.0	90.2	21.6	88.4	51.6	86.8	29.0	4.4	3.0	0.0	0.0	14.8	90.2	52.5	NAO
84	RECIFE	2021	5	12.2	0.0	13.6	42.4	44.4	17.4	40.2	0.6	18.2	19.6	0.0	0.0	14.8	44.4	29.6	NAO

Do conjunto em questão, foram extraídos os valores acumulados de precipitação (chuva) para cada semana dos respectivos meses de cada ano mencionado. Os dados foram agrupados e obteve-se um valor a partir a mediana de todos os valores semanais. O máximo valor de cada semana foi extraído, obtendo assim um valor comparativo para o limiar de acúmulo do volume de chuvas, conforme ilustrado na Figura 2.

De acordo com o Centro Nacional de Monitoramento e Alertas de Desastres Naturais (Cemaden), em apenas 24h (das 11h do dia 23 de maio de 2023 às 11h do dia 24 de maio de 2023), Recife chegou a um acúmulo de 120mm, o suficiente para causar alagamentos de ruas e avenidas. Para o cenário em questão, foram considerados valores de precipitação a partir dos 90mm, caracterizando um volume entre moderado a alto e que poderão ser suficientes para deflagar transtornos diversos. Com isto, este foi o valor de limiar crítico (*threshold*) considerado como "sim" para o rótulo de acúmulo de chuvas na base de dados.

C. Recursos computacionais utilizados

A linguagem de programação utilizada foi o *Python*. Caracteriza-se por ser uma linguagem interativa, interpretada e orientada a objetos, permitindo trabalhar rapidamente e integrar sistemas de forma eficaz [25]. A implementação foi baseada em [10] e se utiliza durante o processo de populares bibliotecas como: *numerical python (NumPy)* que suporta o processamento de arranjos multi-dimensionais e matrizes, juntamente com uma grande coleção de funções matemáticas de alto nível para operar sobre estas matrizes [26]; o *Pandas* é uma ferramenta de análise e manipulação de dados de código aberto rápida, poderosa, flexível e fácil de usar, construída sobre o *Python* [27]; o *Matplotlib* é abrangente para criar visualizações estáticas, animadas e interativas [28]; e a biblioteca de aprendizado de máquina de código aberto *Scikit-learn* [29]. Esta última, por sua vez, suporta aprendizado supervisionado e não supervisionado, fornece várias ferramentas de ajuste de modelo, pré-processamento de dados, seleção de modelo, avaliação, dentre outros utilitários. Sua versatilidade permitiu a comparação entre os resultados das técnicas utilizadas neste trabalho. A execução do código foi realizada por meio da ferramenta online do *Google Colab* também chamada de *Colaboratory* [30] tratando-se de uma plataforma que permite escrever e executar *Python* direto em um navegador Web.

D. Detalhamento dos Experimentos

Os dados utilizados nos experimentos estão contidos em um arquivo (.csv), sendo a última coluna a que informa ao modelo se um acúmulo de chuva ocorreu ou não naquela semana em forma de Sim ou Não. As classificações são feitas com base na ocorrência semanal do volume de chuva naquele ano específico. Para determinar a capacidade de adaptação do modelo, algumas fases foram executadas na base para posteriores experimentos, conforme descrito na Tabela I.

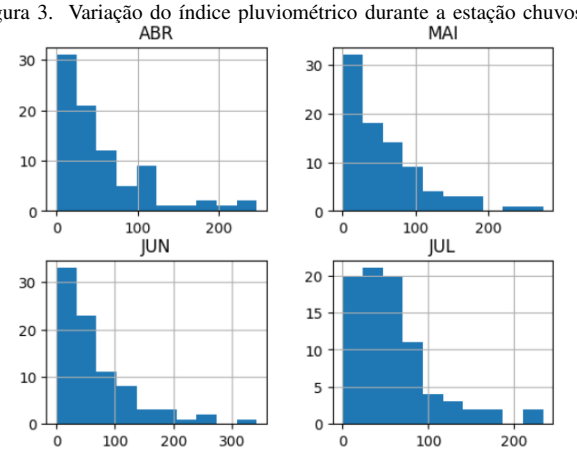
Tabela I
VISÃO GERAL DAS FASES DE PRÉ-PROCESSAMENTO DOS DADOS.

Fase	**Descrição**
Seleção	Seleção dos dados mais representativos.
Verificação	Se há existência de valores nulos.
Transformação	Conversão em determinado tipo de normalização.
Avaliação	Identificação dos padrões mais representativos.
Validação	Treinamento e teste dos dados.

Inicialmente, foi realizado um pré-processamento que consistiu na fase de seleção dos dados mais representativos para o modelo, ou seja, dados de precipitação. Considerando que

o objetivo principal do estudo é a classificação do acumulado de chuva, o fenômeno de precipitação foi priorizado, considerando uma base com 86 linhas e 19 colunas de dados diferentes, indicando os níveis de acúmulo semanal e com uma coluna rotulando se houve ou não o acúmulo de chuva para um determinado período específico. Os dados foram extraídos de uma base pública do INMET [7] e foram escolhidos considerando aspectos de delimitação de escopo e reprodutibilidade do estudo. Na fase seguinte, foi realizada uma verificação da existência de valores nulos, a fim de evitar qualquer disparidade na informação a ser passada para o modelo. Na sequência, os dados descritivos presentes no conjunto são convertidos em formato numérico.

Figura 3. Variação do índice pluviométrico durante a estação chuvosa.

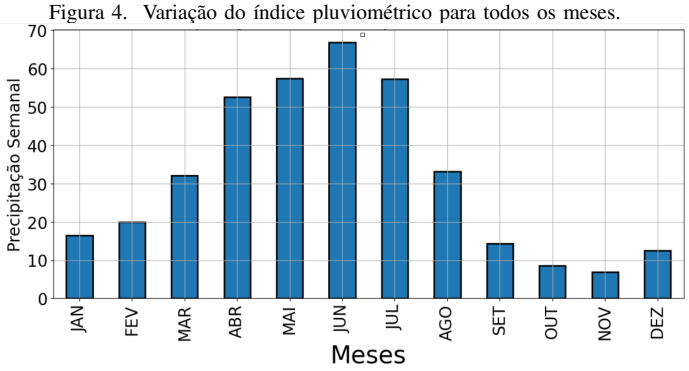


Considerando que os modelos de aprendizado de máquina não funcionam diretamente de forma categórica, tendo como característica para este caso, a sua normalização. Assim que os dados são codificados em formato numérico, o conjunto é avaliado com base na identificação dos padrões mais representativos, sendo por fim, divididos em duas formas: conjunto de treinamento (70%) e conjunto de teste (30%). A partir daqui, a etapa de validação é importante para verificar se o modelo está sendo treinado corretamente com os resultados dos dados de teste. Após o pré-processamento procedeu-se uma breve análise exploratória da base de dados com propósito de avaliar a qualidade, extraindo algumas informações e identificando possíveis relações entre a variável meteorológica selecionada. Assim, durante os experimentos, foi realizada uma análise da variação para os meses com maior estação chuvosa.

A probabilidade de dias com precipitação em Recife varia acentuadamente ao longo do ano. Com base nos dados coletados, a Figura 3 demonstra uma variação durante a estação chuvosa, onde as máximas precipitações semanais para os meses com maior variação são, em valores decrescentes, respectivamente: 3ª semana de junho de 2010 (342,4mm); 1ª semana de maio de 2011 (275,6mm); 2ª semana de abril de 2021 (246,4mm); e 4ª semana de julho de 2019 (234,4mm).

Uma análise do gráfico da Figura 4 retrata os meses com maior e menor precipitação. Verifica-se que a alta quantidade de chuva concentram-se em torno dos meses de abril a julho.

Figura 4. Variação do índice pluviométrico para todos os meses.

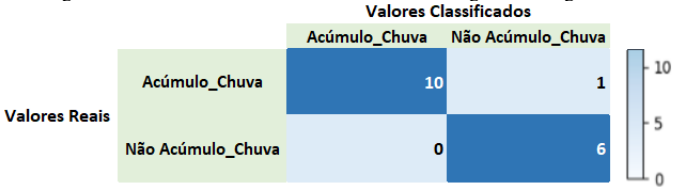


Tal resultado é obtido como média ao longo dos 17 anos de precipitação elencados na base de dados do estudo. O mês de junho têm o maior impacto no modelo e, consequentemente, os meses de setembro a dezembro, contribuem com os menores impactos.

V. Resultados

Este trabalho investigou qual o melhor algoritmo de classificação para identificação de volumes de chuvas críticos, aplicados em dados da cidade do Recife. Foram considerados os algoritmos *k-nearest neighbors (KNN)*, *logistic regression (LR)*, *support vector machine (SVC)*, *decision tree (DT)* e *random forest (RF)*. O desempenho dos algoritmos foi avaliado de acordo com *accuracy*, *recall* e *ROC*, identificando quão preciso são os algoritmos utilizados em termos de classificação.

Figura 5. Matriz de Confusão do modelo de Regressão Logística.



A Figura 5 ilustra a matriz de confusão para o modelo de *logistic regression (LR)*. Uma matriz de confusão é uma tabela que permite a visualização do desempenho de um algoritmo de classificação. Ao analisar a matriz de confusão do modelo *logistic regression* representado na Figura 5, verifica-se que o algoritmo classificou corretamente o acúmulo de chuva em dez vezes e o não acúmulo em seis vezes; assim como classificou incorretamente o acúmulo de chuva em uma vez e incorretamente o não acúmulo em zero vezes. Com isto, conclui-se que o modelo classificou acertivamente 16 das 17 classificações, o que carecterizou uma *Accuracy Score* de 94,12%, um *Recall Score* de 100% e um *ROC Score* de 95,45%, conforme sumarizado na Tabela II. Esta tabela resume o desempenho de todos os algoritmos considerados no estudo.

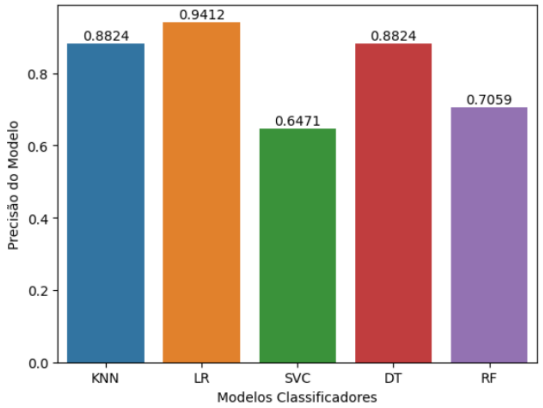
Percebe-se que a *logistic regression (LR)* apresentou os melhores resultados em comparação com os demais algoritmos em termos de acurácia. Além disso, usando LR, as chances de classificar falsamente um valor positivo são menores em comparação com as demais técnicas.

Tabela II
RESULTADOS DOS MODELOS DE CLASSIFICAÇÃO DO CONJUNTO DE TESTE.

Modelo	Accuracy Score	Recall Score	ROC Score
KNN	88,24	66,67	83,34
LR	**94,12**	**100**	**95,45**
SVC	64,71	0	50
DT	88,24	83,34	87,12
RF	70,59	83,34	87,12

Observa-se ainda que há para o modelo obtido por LR um desempenho muito superior em relação as demais métricas avaliadas, o que permite concluir que a *logistic regression* possui um conceito muito bom na classificação.

Figura 6. Comparativo de resultados entre os 5 modelos estudados.



Ainda em termos de acurácia, os modelos *KNN* e *DT*, em comparação, tiveram o mesmo valor superior a 88% e a *RF* um valor superior a 70%, também mostrando-se abaixo da *LR* em destaque. Bem ao final, a *SVC* aparece com valor superior a 64% não tendo demonstrado a mesma eficiência percebida em outros estudos na literatura. Os valores de acurácia da Tabela II estão representados graficamente na Figura 6.

A partir do comparativo de resultados, verificou-se com clareza o destaque da *logistic regression (LR)* ao superar os demais modelos de aprendizado de máquina. Uma das vantagens deste modelo é que é preciso saber apenas se um evento (ocorrência ou não de acúmulo de chuva, por exemplo) ocorreu para então utilizar um valor dicotômico, ou seja, aquele em que a classificação cujas divisões ou partes apresentam somente dois termos, como variável dependente. A partir deste valor, o procedimento classifica sua estimativa de que houve ou não, neste caso, o acúmulo do volume de chuva. Com isto, cada elemento é classificado de acordo com a maior probabilidade prevista de pertencer a um grupo. Pelos resultados observados, recomenda-se o modelo de LR para um processo de classificação para o problema e cenário adotado neste estudo.

VI. CONCLUSÃO

Este trabalho investigou qual o melhor algoritmo de classificação para identificação de volumes de chuvas críticos, aplicados em dados da cidade do Recife. Foram considerados os algoritmos *KNN*, *LR*, *SVC*, *DT* e *RF*. O desempenho dos algoritmos foi avaliado de acordo com *accuracy*, *recall* e *ROC*, identificando quão preciso são os algoritmos utilizados em termos de classificação. Dentre os algoritmos explorados, a *logistic regression (LR)* destacou-se em todas as métricas avaliadas e em comparação aos demais. Desta forma, a técnica LR poderia ser adotada como algoritmo preferencial em um futuro sistema de emissão de alertas de acumulado de chuvas para o cenário estudado.

Trabalhos futuros podem investigar a integração do motor de classificação inteligente com um sistema para emissão de alertas de eventos críticos decorrentes de precipitações intensas. Além disso, o estudo pode ser expandido futuramente para considerar outras fontes de dados e o cruzamento destes dados, incluindo: outras bases (como a base do DECEA), dados de estações metereológicas próprias, dados de satélite, assim por diante. Outra investigação importante está relacionada com intervalos temporais de mais curto prazo, de 48 até 72 horas.

AGRADECIMENTOS

Os autores agradecem o apoio financeiro fornecido pela Fundação de Amparo a Ciência e Tecnologia de Pernambuco - (FACEPE) e Universidade Federal Rural de Pernambuco - (UFRPE).

REFERÊNCIAS

[1] Kuwajima, Julio Issao and Fan, Fernando Mainardi and Schwanenberg, Dirk and Assis Dos Reis, Alberto and Niemann, André and Mauad, Frederico Fábio. Climate change, water-related disasters, flood control and rainfall forecasting: a case study of the São Francisco River, Brazil. Geological Society, London, Special Publications. 2019.

[2] IPCC - Intergovernmental panel on climate change, 2022: climate change 2022 – impacts, adaptation and vulnerability. Disponível em: https://ipcc.ch/. Acesso em: 13/11/2022.

[3] Recife (2019). Análise de riscos e vulnerabilidades climáticas e estratégia de adaptação do município do recife. Disponível em: https://www2.recife.pe.gov.br/sites/default/files/sumario-clima-recife-portugues.pdf. Acesso em: 15/11/2022.

[4] ALMEIDA, L.; SERRA, J. C. V. Modelos hidrológicos, tipos e aplicações mais utilizadas. Revista da FAE, v. 20, n. 1, p. 129–137, 2017.

[5] DUNCAN, A.; CHEN, A. S.; KEEDWELL, E.; DJORDJEVIC, S.; SAVIC, D. Urban flood prediction in real-time from weather radar and rainfall data using artificial neural networks. International Association of Hydrological Sciences, 2011.

[6] Drosdek. J, Brozoski. R. L. Aprimoramento na Previsão de Inundações e Enchentes com Aprendizado de Máquina. Ed. da UnC, 2021.

[7] INMET. Instituto Nacional de Meteorologia. Disponível em: https://tempo.inmet.gov.br/TabelaEstacoes/A301. Acesso em: 15/11/2022.

[8] Darji, M. P., Dabhi, V. K., and Prajapati, H. B. (2015). Rainfall forecasting using neural network: A survey. In International Conference on Advances in Computer Engineering and Applications, India. IMS Engineering College.

[9] Alpaydin, E. (2010). Introduction to Machine Learning. The MIT Press, 2 edition.

[10] Sai Prasanth Kadiyala and Wai Lok Woo. 2022. Flood Prediction and Analysis on the Relevance of Features using Explainable Artificial Intelligence. In 2021 2nd Artificial Intelligence and Complex Systems Conference (AICSconf '21). Association for Comuting Machinery, New York, NY, USA, 1-6.

[11] IBM. IA explicável. Disponível em: https://www.ibm.com/br-pt/watson/explainable-ai. Acesso em: 01/06/2023.

[12] N. Tengtrairat, W.L. Woo, P. Parathai, C. Aryupong, P Jitsangiam, D. Rinchumphu, in: Automated Landslide-Risk Prediction using Web GIS and Machine Learning Models. Sensors 2021, 21(13), 4620.

[13] B.H.D. Koh, C.L.P. Lim, H. Rahimi, W.L. Woo, B. Gao, in: Deep Temporal Convolution Network for Time Series Classification. Sensors 2021, 21(2), 603.
[14] W.L. Woo, in: Human-Machine Co-Creation in the Rise of AI. IEEE Instrumentation and Measurement Magazine, 23(2), pp. 71-73, (2020).
[15] Facco, M., Campos, M. A. de A. de, Vargas, D. dos S., Silveira, R. B., Bisognin, C. (2020). Algoritmos de Machine Learning Aplicados na Ocorrência de Chuvas na Cidade de Santa Maria. Ciência E Natura, 42.
[16] SOUSA, Rafaela dos Santos; GUEDES, Elloá B.; OLIVEIRA, Maria Betânia Leal de. Previsão Anual de Precipitações em Manaus, Amazonas: Um comparativo de técnicas de Aprendizado de Máquina. In: WORKSHOP DE COMPUTAÇÃO APLICADA À GESTÃO DO MEIO AMBIENTE E RECURSOS NATURAIS (WCAMA), 2018, Natal. Anais. Porto Alegre: SBC, 2018.
[17] Radio Agência Nacional, 2023. Volume de chuva em Recife chega a 130% do previsto para fevereiro. Disponível em: https://agenciabrasil.ebc.com.br/radioagencia-nacional/geral/audio/2023-02/volume-de-chuva-em-recife-chega-130-do-previsto-para-fevereiro. Acesso em: 22/05/2023.
[18] Folha de Pernambuco, 2022. Acumulado de chuvas das últimas 24h no Recife é o terceiro maior dos últimos 50 anos. Disponível em: https://www.folhape.com.br/noticias/acumulado-de-chuvas-das-ultimas-24h-no-recife-e-um-dos-maiores-dos/228053/. Acesso em: 22/05/2023.
[19] Sambatti, S. B. M. ; Martins, R. G. ; Vilela, R. B. ; Cotacallapa, F. M. ; Pessoa, A. S. A. ; Dias, J. ; Bressiani, D. A. ; Fernandes, G. P. Previsão de riscos de alagamentos e inundações com uso de inteligência artificial. Revista de Informática Aplicada, 2019.
[20] Fernandes, A. A. T., Figueiredo Filho, D. B., Rocha, E. C. da., Nascimento, W. da S. 2020. Leia este artigo se você quiser aprender regressão logística. Revista de Sociologia e Política, v. 28, n.74.
[21] Santos, E. J. R. Previsão de precipitação usando máquinas de vetores de suporte visando sua implementação em sistemas embarcados. Dissertação de Mestrado. Universidade de Brasília. Brasília, 2019.
[22] Witten, I. H. and Frank, E. Data mining: practical machine learning tools and techniques.Morgan Kaufmann Publishers, San Francisco. 2005.
[23] Meira, C. A. A. Processo De Descoberta De Conhecimento Em Bases De Dados Para a Analise E O Alerta De Doenças De Culturas Agricolas E Sua Aplicação Na Ferrugem Do Cafeeiro. 2008.
[24] Brito, L. A. V. Modelo de classificação multivariável para identificação de enchentes: um estudo empírico no sistema de monitoramento de rios e-noe. 2019. Dissertação (Mestrado em Ciências de Computação e Matemática Computacional) – ICMC, Univ. de São Paulo, São Carlos, 2019.
[25] Python. Disponível em: https://www.python.org/. Acesso em: 05/11/2022.
[26] NumPy. Disponível em: https://numpy.org/. Acesso em: 07/11/2022.
[27] Pandas. Disponível em: https://pandas.pydata.org/. Acesso em: 07/11/2022.
[28] MatPlotLib. Disponível em: https://matplotlib.org/. Acesso em: 07/11/2022.
[29] Scikit-learn. Machine Learning in Python. Disponível em: https://scikit-learn.org/stable/. Acesso em: 07/11/2022.
[30] Colab Research. Disponível em: https://colab.research.google.com/. Acesso em: 05/11/2022.
[31] Rodrigues, V. Entenda o que é AUC e ROC nos modelos de Machine Learning. Disponível em: https://medium.com/towards-data-science/understanding-auc-roc-curve-68b2303cc9c5. Acesso em: 27/05/2023.