UNIVERSIDADE FEDERAL DE SANTA MARIA
CENTRO DE TECNOLOGIA
CURSO DE GRADUAÇÃO EM CIÊNCIA DA COMPUTAÇÃO

Andriza Maria da Cunha Campanhol

EXPLICABILIDADE DE MODELO LONG SHORT-TERM MEMORY (LSTM) PARA PREVISÃO DE DENGUE EM PORTO ALEGRE

Santa Maria, RS
2024

Andriza Maria da Cunha Campanhol

EXPLICABILIDADE DE MODELO LONG SHORT-TERM MEMORY (LSTM) PARA PREVISÃO DE DENGUE EM PORTO ALEGRE

Trabalho Final de Graduação apresentado ao Curso de Graduação em Ciência da Computação da Universidade Federal de Santa Maria (UFSM, RS), como requisito parcial para obtenção do grau de **Bacharel em Ciência da Computação**.

Orientador: Prof. Joaquim Vinicius Carvalho Assunção

Santa Maria, RS
2024

Andriza Maria da Cunha Campanhol

EXPLICABILIDADE DE MODELO LONG SHORT-TERM MEMORY (LSTM) PARA PREVISÃO DE DENGUE EM PORTO ALEGRE

Trabalho Final de Graduação apresentado ao Curso de Graduação em Ciência da Computação da Universidade Federal de Santa Maria (UFSM, RS), como requisito parcial para obtenção do grau de **Bacharel em Ciência da Computação**.

Aprovado em 30 de julho de 2024:

_____

Joaquim Vinicius Carvalho Assunção, Dr. (UFSM)
(Presidente/Orientador)

_____

Leonardo Ramos Emmendorfer, Dr. (UFSM)

_____

Daniel Fernando Tello Gamarra, Dr. (UFSM)

Santa Maria, RS
2024

RESUMO

EXPLICABILIDADE DE MODELO LONG SHORT-TERM MEMORY (LSTM) PARA PREVISÃO DE DENGUE EM PORTO ALEGRE

AUTORA: Andriza Maria da Cunha Campanhol
Orientador: Joaquim Vinicius Carvalho Assunção

A dengue é um crescente problema de saúde no Brasil nos últimos anos, especialmente em regiões como Porto Alegre, onde sua incidência tem sido significativa (SMS-POA, 2024). Caracterizada por um padrão sazonal e fortemente influenciada por fatores climáticos, a dengue apresenta desafios significativos no que diz respeito à previsão e controle de sua propagação (GOV-MS, 2024). Embora métodos de forecasting tenham sido amplamente utilizados para predição da doença, a falta de explicabilidade desses modelos os torna menos claros sobre os fatores que podem influenciar suas previsões. Dessa forma, este trabalho busca explicar a saída de modelos preditivos de valores contínuos, utilizando dados temporais relativos as condições climáticas e casos de dengue em Porto Alegre. O objetivo é abordar problemas de Explicabilidade de Inteligência Artificial (XAI) em modelos Long Short-Term Memory (LSTM) para previsão, empregando técnicas de Symbolic Aggregate approXimation (SAX) e apresentando os resultados de forma explicável através de Árvores de Decisão.

Palavras-chave: LSTM. XAI. Dengue.

ABSTRACT

EXPLAINABILITY OF AN LONG SHORT-TERM MEMORY (LSTM) MODEL FOR DENGUE FORECASTING IN PORTO ALEGRE

AUTHOR: Andriza Maria da Cunha Campanhol
ADVISOR: Joaquim Vinicius Carvalho Assunção

Dengue has become a growing health problem in Brazil in recent years, especially in regions like Porto Alegre, where its incidence has been significant (SMS-POA, 2024). Characterized by a seasonal pattern and strongly influenced by climatic factors, dengue presents significant challenges in terms of prediction and control of its spread (GOV-MS, 2024). Although forecasting methods have been widely used for prediction, the lack of explainability of these models makes it less clear what factors may influence their predictions. Therefore, this work seeks to explain the output of predictive models for continuous values, using temporal data related to climatic conditions and cases of dengue in Porto Alegre. The objective is to address Explainable Artificial Intelligence (XAI) issues in Long Short-Term Memory (LSTM) models for forecasting, employing Symbolic Aggregate approXimation (SAX) techniques and presenting the results in an understandable way through Decision Trees.

Keywords: LSTM. XAI. Dengue.

LISTA DE FIGURAS

LISTA DE TABELAS

LISTA DE ABREVIATURAS

ANN	Artificial Neural Network
CNN	Convolutional Neural Network
DNN	Deep Neural Network
dwSAX	Distribution-Wise Symbolic Aggregate approXimation
LSTM	Long Short-Term Memory
MAE	Mean Absolute Error
PAA	Piecewise Aggregate Approximation
qSAX	Quantile Symbolic Aggregate approXimation
RNN	Recurrent Neural Network
SAX	Symbolic Aggregate approXimation
SVM	Support Vector Machine
XAI	eXplainable AI

SUMÁRIO

1 INTRODUÇÃO

A dengue é uma doença que pertence ao grupo das arboviroses, que são causadas por vírus transmitidos por insetos artrópodes. No Brasil, o vetor responsável pela transmissão da dengue é o mosquito *Aedes aegypti*. O vírus da dengue (DENV) é transmitido ao ser humano principalmente pela picada de fêmeas infectadas desse mosquito. A primeira epidemia de dengue registrada clinicamente e confirmada por testes laboratoriais no Brasil ocorreu entre 1981 e 1982, na cidade de Boa Vista, em Roraima (GOV-MS, 2024).

Nos últimos anos, houve um aumento significativo no número de casos de dengue no Brasil, o que é preocupante, já que, embora a maioria dos pacientes se recupere, alguns podem desenvolver formas graves da doença, inclusive virem a óbito. Assim, a qualidade da assistência médica e a organização dos serviços de saúde desempenham um papel crucial nesse cenário. No estado do Rio Grande do Sul não é diferente, a cidade de Porto Alegre já registra mais de 1.800 casos de dengue no ano de 2024 até a metade de abril (SMS-POA, 2024).

1.1 JUSTIFICATIVA

A urbanização desordenada, o crescimento populacional acelerado, a falta de saneamento básico e as condições climáticas contribuem para manter um ambiente propício à proliferação do vetor, afetando a transmissão dos arbovírus, incluindo o da dengue. A dengue apresenta um padrão sazonal, com um aumento significativo de casos e risco de epidemias, principalmente entre outubro e maio do ano seguinte (GOV-MS, 2024).

Nesse contexto, os algoritmos de mineração de dados apresentam ser uma boa alternativa para investigar e descobrir novas informações em grandes quantidades de dados relacionados a doença. Esses algoritmos são capazes de identificar padrões complexos e relações não lineares nos dados, possibilitando a descoberta de novos conhecimentos, não facilmente identificados por métodos tradicionais.

O *forecasting* é um método que utiliza de informações anteriores, por meio de técnicas estatísticas, para realizar previsões ou projeções futuras. Dentre os modelos computacionais utilizados com frequência estão a regressão linear, as *Support Vector Machines* (SVMs) e as *Artificial Neural Networks* (ANNs). Embora existam várias técnicas para isso, as *Recurrent Neural Network* (RNNs) e as *Long Short-Term Memory* (LSTMs), devido à sua capacidade de capturar e aprender com padrões sequenciais complexos nos dados, são altamente eficazes em previsões de séries temporais. Em especial, as LSTMs resolvem o problema de esquecimento temporal (ver Seção 2.2).

Porém, esses métodos não são passíveis de apresentar uma explicação de seus resultados, pois suas operações internas (como os estados de células e portões que regulam o fluxo de informações) não são facilmente interpretáveis. Assim, somente o resultado de inteligências artificiais não explicativas, ou caixas-pretas, acabam não sendo confiáveis para diagnósticos e previsões de dados da saúde.

1.2 MOTIVAÇÃO

A área de *eXplainable AI* (XAI), vem sendo amplamente explorada na literatura, buscando tornar as inteligências artificiais mais transparentes e compreensíveis, como nos trabalhos recentes de (AHMED et al., 2023) e (SOROUSH; RAJI; GHAVAMI, 2023). No entanto, especialmente no contexto da predição da dengue, a maior parte dos estudos concentra-se em técnicas de regressão ou não apresenta explicabilidade (Ver Capítulo 3).

Nesse contexto, o presente trabalho visa explicar a saída de modelos preditivos para valores numéricos contínuos, utilizando os dados temporais das condições climáticas e casos de dengue na cidade de Porto Alegre. Para alcançar esse objetivo, são utilizadas redes neurais LSTM juntamente com a *Symbolic Aggregate approXimation* (SAX), a fim de classificar os dados em diferentes grupos e relacioná-los com as diversas variáveis presentes no banco de dados. Também sendo empregadas as técnicas de *Random Forest* e *Decision Tree* para representar as correlações identificadas (ver Capítulo 4).

Portanto, o processo proposto consegue resolver problemas relacionados a XAI para *forecasting* em execuções de modelos LSTMs, utilizando técnicas de SAX, obtendo padrões entre diferentes variáveis relacionadas e mostrando o resultado de forma explicável via árvores de decisão.

1.3 OBJETIVOS

1.3.1 Objetivo geral

Este trabalho busca propor um modelo para resolver problemas de XAI na predição de dengue em modelos LSTM, utilizando dados temporais das condições climáticas e casos de dengue em Porto Alegre.

1.3.2 Objetivos específicos

- Implementar um modelo preditivo utilizando uma rede neural LSTM para prever os casos de dengue referentes ao ano de 2024.
- Utilizar a técnica de SAX, com variações, para classificar os dados temporais e preditos em símbolos, a fim de identificar os picos de casos e correlações com os dados climáticos, para duas e cinco classes.
- Aplicar o algoritmo de *Random Forest* para descobrir as variáveis mais influentes nas explosões de casos de dengue classificadas.
- Empregar árvores de decisão para criar representações explicativas dos resultados do modelo, baseadas nas características climáticas encontradas, buscando as de maior precisão.

1.4 ORGANIZAÇÃO DO TRABALHO

O trabalho está organizado da seguinte forma: no Capítulo 1, é apresentada a introdução, que contextualiza o problema da dengue, sua relevância e os objetivos da pesquisa. No Capítulo 2, é apresentado o referencial teórico, abordando as técnicas de mineração de dados e aprendizado de máquina aplicadas no estudo. O Capítulo 3 apresenta uma breve revisão dos trabalhos relacionados. O Capítulo 4 descreve a metodologia utilizada, detalhando a organização e transformação dos dados, as técnicas de análise de dados e os algoritmos de aprendizado de máquina empregados. No Capítulo 5, são discutidos os resultados e suas análise. Por fim, o Capítulo 6 apresenta as conclusões e possíveis trabalhos futuros.

2 REFERENCIAL TEÓRICO

Nesse capítulo, estão apresentadas as principais técnicas e tecnologias utilizadas para o desenvolvimento deste trabalho, tais como XAI, LSTM, SAX, Árvore de Decisão e *Random Forest*.

2.1 EXPLAINABLE AI (XAI)

A XAI é uma área de pesquisa que busca utilizar métodos e conjuntos de processos que tornam as saídas e resultados dos modelos de IA mais compreensíveis, permitindo que os usuários possam entender o funcionamento e confiar nas decisões tomadas pelas inteligências artificiais. Portanto, aplicar esses modelos em contextos sensíveis e críticos para a segurança, como na área da saúde e no campo jurídico, onde a transparência e a confiabilidade nas decisões são essenciais, se torna especialmente desafiador.

Esses métodos são necessários, pois as operações internas, como células e portões que controlam o fluxo de informações nas redes neurais, não são facilmente interpretáveis pelos humanos. Os algoritmos de inteligência artificial, a medida em que crescem suas complexidades de compreensão, são comumente chamados de caixas-pretas, pois não são capazes de apresentar uma explicação dos seus resultados sem utilizar técnicas auxiliares como as de XAI.

Muitos trabalhos recentes estão aplicando os métodos de XAI para diversos tipos de problemas, como em (AHMED et al., 2023) onde são utilizados conjuntamente com técnicas de aprendizado de máquina para prever casos de Diabetes tipo 2. Além disso, (SOROUSH; RAJI; GHAVAMI, 2023), empregaram XAI para compreender o funcionamento das DNNs com o objetivo de realizar a compressão desses modelos.

Essas técnicas estão sendo exploradas principalmente no intuito de validar a precisão de previsões e em tarefas de classificação, especialmente quando implementadas para interpretação de DNNs, SVMs e ANNs (DOŠILOVIĆ; BRČIĆ; HLUPIĆ, 2018). A escolha da técnica adequada depende do tipo de modelo utilizado e do que se deseja compreender.

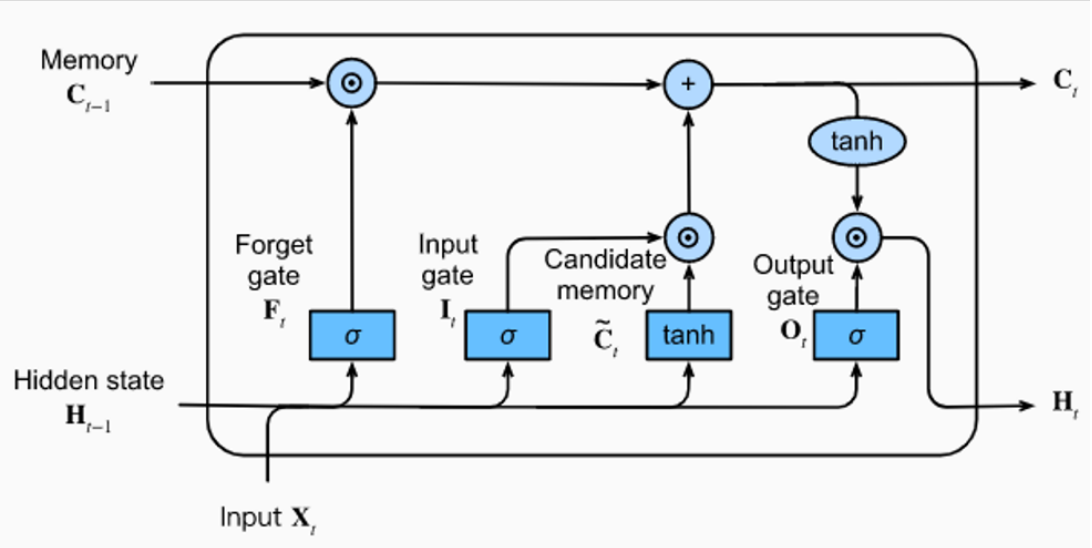
Existem diferentes tipos de explicadores em XAI. Explicadores intrínsecos são modelos que são interpretáveis por design, como árvores de decisão. Explicadores pós-hoc são funções que fornecem explicações para as decisões de modelos de caixa-preta, sendo independentes do modelo original. Eles podem ser globais, oferecendo explicações para todas as decisões possíveis, ou locais, explicando apenas decisões específicas. Além disso, explicadores podem ser independentes, funcionando com qualquer tipo de modelo, ou específicos, dependendo da estrutura do modelo original (GUIDOTTI et al., 2021).

2.2 LONG SHORT-TERM MEMORY (LSTM)

A LSTM é uma variação das *Recurrent Neural Network* (RNNs) que foi proposta por (HOCHREITER; SCHMIDHUBER, 1997) para abordar problemas como desaparecimento e explosão do gradiente. Diferentemente das RNNs convencionais, esse modelo é capaz de aprender relações de longo e curto prazo, reduzindo a dependência excessiva de informações de curtíssimo prazo.

Na arquitetura de uma RNN, ela basicamente consiste em uma única camada de neurônios e se retroalimenta com a mesma. Já na arquitetura da LSTM, apresentada na Figura 1, há um fluxo mais complexo de informações, permitindo o armazenamento e processamento de séries temporais ao longo dos módulos. Ela é comporta por células de memória, cada uma formada por unidades de memória e portões lógicos, que controlam esse fluxo de informações na rede através de funções de sigmoide e tanh, chamadas de funções de ativação.

Figura 1 – Arquitetura de uma célula em um modelo LSTM.



Fonte: Memória Longa de Curto Prazo (LSTM) - Dive into Deep Learning

Cada saída da função de sigmoide funciona como um filtro que define a quantidade de informação que irá passar, variando de 0 a 1, que é um peso multiplicado na sequência pelas entradas. Quando uma saída é igual a 1, significa que toda informação da entrada irá passar, enquanto uma saída de 0 indica que nenhuma informação irá passar.

A estrutura da LSTM apresenta 3 portões lógicos que utilizam as funções de ativação. O *forgot gate* (portão de esquecimento) recebe duas entradas Xt (entrada atual) e Ht-1 (saída da célula anterior), essas entradas passam pelo portão, onde são multiplicadas por matrizes de peso e adicionadas de um um viés. O resultado passa então pela função de ativação, que elimina as informações que não são mais úteis.

O *input gate* (portão de entrada) regula a informação das entradas Xt e Ht-1 utilizando a função de sigmoide e cria um vetor através da função tanh, contendo todos os valor possíveis das entradas. Assim, é feita a multiplicação dos valores regulados e do vetor para obter as informações úteis a serem adicionadas ao estado da célula.

Por fim, o *output gate* (portão de saída) inicia utilizando a função tanh na célula e regulando os valores das entradas Xt e Ht-1 com a função de sigmoide para filtrar os que devem ser lembrados. O resultado é obtido pela multiplicação do vetor e dos valores regulados, para extrair as informações úteis do estado da células atual e servir de entrada para a próxima célula.

Abaixo estão apresentadas as formulações matemáticas das células na LSTM:

$$F_t = \sigma(W_f X_t + U_f H_{t-1} + b_f) \tag{2.1}$$

$$I_t = \sigma(W_i X_t + U_i H_{t-1} + b_i) \tag{2.2}$$

$$O_t = \sigma(W_o X_t + U_o H_{t-1} + V_o C_t + b_o) \tag{2.3}$$

$$C_t = f_t \odot C_{t-1} + i_t \odot \tilde{C}_t \tag{2.4}$$

$$H_t = O_t \odot \text{tanh}(C_t) \tag{2.5}$$

As equações 2.1, 2.2 e 2.3 representam os portões de esquecimento, entrada e saída, respectivamente, sendo F_t, I_t e O_t os vetores de ativação correspondentes. A equação 2.4 representa a atualização do estado da célula e a equação 2.5 representa a saída da LSTM. Nessas equações, X_t é o vetor de entrada, H_t é o vetor de saída, σ representa a função sigmoid e t o tempo atual. W, U e b são matrizes de pesos e parâmetros do vetor de viés que são aprendidos durante o treinamento.

2.3 SYMBOLIC AGGREGATE APPROXIMATION (SAX)

A SAX é uma técnica de representação simbólica de séries temporais, sua proposta se torna adequada na resolução dos principais problemas encontrados nos modelos anteriores: a alta dimensionalidade da representação simbólica e a correlação das medidas de distância com os dados originais. O artigo de (LIN et al., 2007) apresenta o modelo e as aplicações da representação em várias tarefas de mineração de dados de agrupamento, classificação, detecção de anomalias, descoberta de padrões e visualização.

O objetivo da SAX é dividir séries temporais contínuas em intervalos e, em seguida, transforma cada série em uma sequência de símbolos, normalmente representados por letras, facilitando o armazenamento, manipulação e comparação dos dados. Para isso, utiliza de duas etapas principais de transformação: a *Piecewise Aggregate Approximation* (PAA), que consiste na redução da dimensionalidade da série temporal, e na sequência a discretização.

Na etapa de PAA, dada a série original X de n pontos, ela é primeiramente dividida em w segmentos de mesmo tamanho, onde $w < n$. Em seguida, é calculado o valor médio de cada segmento, criando uma nova série que representa a original de maneira mais compacta. Assim, a representação preserva as principais tendencias da série temporal, diminuindo a quantidade de dados sem perder muita informação.

A fórmula 2.6 mostra o cálculo do valor médio de cada segmento, onde i representa o índice do segmento, j representa o índice dos pontos dentro do segmento, $\bar{X}_i$ é o valor médio do i-ésimo segmento de PAA, X_j representa os pontos individuais na série temporal original, e a soma se estende do início ao fim de cada segmento. Cada valor $\bar{X}_i$ da série PAA representa, portanto, a média dos pontos dentro de seu segmento, fornecendo uma representação simplificada e de menor dimensionalidade da série original.

$$\bar{X}_i = \frac{1}{w} \sum_{j=\frac{n}{w}(i-1)+1}^{\frac{n}{w}\cdot i} X_j \tag{2.6}$$

Já na etapa de discretização, a série temporal simplificada é transformada em uma sequência de símbolos. Para isso, são estabelecidos intervalos de valores baseados na distribuição normal padrão e são atribuídos símbolos específicas para cada intervalo. Esses intervalos são definidos pelos *breakpoints*, que são são determinados previamente e dependem do tamanho do alfabeto escolhido para a representação simbólica. Na sequência, os valores calculados na etapa de PAA, representados por $\bar{X}_i$, são mapeados para os símbolos correspondes aos intervalos aos quais pertencem em relação aos *breakpoints*, concluindo assim o processo de transformação.

Um dos problemas enfrentados pela SAX é a garantia do balanceamento das classes, pois assume que os dados seguem uma distribuição normal (Gaussiana). Com essa suposição, cada intervalo tem a mesma área sobre a curva de distribuição, garantindo assim que cada símbolo tenha a mesma probabilidade de ocorrência. Assim, quando os dados não seguem a distribuição normal, pode levar a um desbalanceamento na representação simbólica. Isso significa que alguns símbolos podem ser mais comuns que outros, podendo afetar o aprendizados dos algoritmos de aprendizado de máquina, favorecendo as classes mais comuns e não conseguindo aprender bem com classes menos comuns.

Diferentes variações da abordagem padrão foram propostas para tentar resolver o problema de balanceamento, abaixo estão descritas duas delas.

2.3.1 Distribution-Wise Symbolic Aggregate approXimation (dwSAX)

Quando a distribuição da série temporal não é gaussiana ou se distorce ao longo do tempo, a SAX pode não fornece uma representação simbólica adequada. Nesse contexto, o artigo (KLOSKA; ROZINAJOVA, 2020) propõe uma nova técnica de representação chamada dwSAX. Essa técnica amplia a abordagem da SAX levando em conta a distribuição dos dados dentro de cada segmento da série temporal, ao invés de se basear apenas no valor médio. Assim, utilizando uma abordagem mais geral para a seleção de pontos de quebra simbólicos, permite que capture características mais complexas das séries temporais.

A dwSAX utiliza o método de Estimativa de Densidade Kernel (KDE) para estimar a distribuição de probabilidade de uma variável aleatória, através de uma amostra de dados independentes e identicamente distribuídos. O KDE utiliza a função de kernel e um parâmetro de suavização para mapear a densidade desconhecida dos dados. Portanto, sua precisão depende do parâmetro de suavização, que pode resultar em uma estimativa distorcida da distribuição dos dados e problemas na análise de intervalos específicos, se não for escolhido adequadamente.

2.3.2 Quantile Symbolic Aggregate approXimation (qSAX)

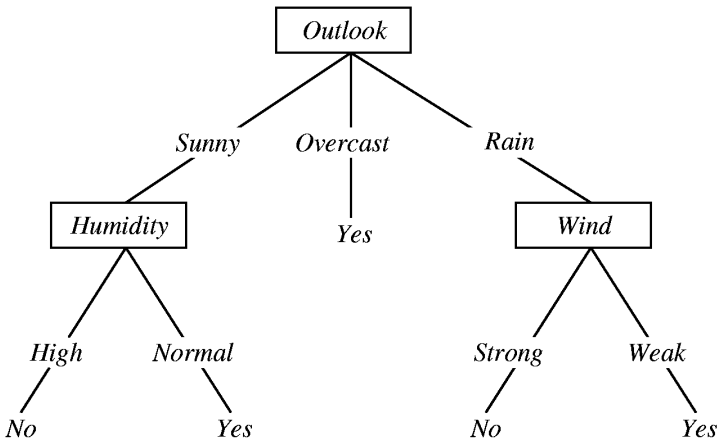
A técnica de qSAX, apresentada em (SILVEIRA; ASSUNÇÃO; EMMENDORFER, 2023), é uma outra variação específica do método SAX, diferenciando-se principalmente na maneira como os pontos de divisão são determinados. Para identificar esses pontos no qSAX, é necessário calcular os quantis dos dados, que são valores que dividem os dados em intervalos de igual probabilidade. Esses pontos de divisão são então usados para transformar a série temporal em uma sequência de símbolos, também conhecidos como classes.

Dessa forma, o qSAX propõe uma abordagem que garante o balanceamento de classes, independentemente da distribuição dos dados de entrada. Esse método garante um balanceamento de classes preciso, ao contrário do dwSAX, que pode apresentar imprecisões devido à sua estimativa da função de densidade de probabilidade dos dados. Portanto, o qSAX produz uma saída equilibrada, independentemente do número de símbolos, o que melhora a consistência e a representatividade dos dados classificados, proporcionando uma análise mais precisa dos padrões temporais.

2.4 ÁRVORE DE DECISÃO

A Árvore de Decisão (QUINLAN, 1986) é um algoritmo supervisionado utilizado principalmente para problemas de regressão e classificação. Algoritmos supervisionados treinam modelos usando dados rotulados, onde cada exemplo de treinamento possui entradas e saídas desejadas. A construção de uma árvore de decisão, chamada indução de árvores de decisão, envolve dividir os dados em subconjuntos com base nos atributos mais importantes até que todas as divisões sejam homogêneas ou atinjam um critério específico. A estrutura da árvore é composta por um nó raiz, nós intermediários e nós folha, com cada nó representando uma característica ou atributo dos dados e cada ramificação representando uma decisão baseada nesse atributo. A representação simples de uma árvore de decisão é apresentado na Figura 2.

Figura 2 – Representação de uma árvore de decisão simples.



Fonte: (QUINLAN, 1986)

Seu objetivo é dividir o conjunto de dados em subconjuntos cada vez menores, a partir das regras de decisão baseadas nos valores dos atributos. Começando pelo nó raiz, a árvore é construída recursivamente até que todos, ou a maioria, dos dados tenham sido classificados em classes específicas, ou tenha atendido algum critério de parada, como profundidade máxima. Para decidir como dividir os dados em cada nó, a árvore de decisão avalia diferentes critérios, como ganho de informação ou índice de Gini, que medem a pureza dos subconjuntos resultantes após a divisão.

Conforme uma árvore de decisão fica maior, é mais difícil manter a pureza dos nós, muitas vezes resultando em subconjuntos de dados muito pequenos e específicos. Isso é chamado de fragmentação de dados e pode levar ao *overfitting*. Por isso, as árvores de decisão tendem a favorecer tamanhos menores, buscando a simplicidade para evitar problemas de *overfitting*, empregando técnicas como a poda para reduzir sua complexidade.

Um dos algoritmos mais utilizados é o CART (*Classification and Regression Trees*), uma abordagem mais geral para árvores de decisão que pode ser usada tanto para problemas de classificação quanto de regressão (BREIMAN et al., 1984).

As árvores de decisão em geral são de fácil interpretação, requerem pouca preparação dos dados e são versáteis para diversos problemas de mineração de dados. Entretanto, é preciso levar em consideração as complicações do *overfitting* e a dificuldade de representação de problemas mais complexos.

2.5 RANDOM FOREST

O *Random Forest* (BREIMAN, 2001) é um algoritmo de aprendizado de máquina também utilizado principalmente em problemas de regressão e classificação. Seu objetivo é combinar a saída de diversas árvores de decisão para obter um único resultado. Para isso, utiliza o método de *Ensemble Learning*, mais especificamente a técnica de *Bagging* (BREIMAN, 1996).

A técnica de *Bagging* é utilizada para criar várias amostas de treinamento, selecionando aleatoriamente amostras com substituição do conjunto de dados originais. Isso significa que uma amostra específica pode aparecer várias vezes na mesma amostra de treinamento. Para cada amostra de treinamento, uma árvore de decisão é construída usando um processo semelhante ao da árvore de decisão tradicional. Cada árvore é treinada com uma parte dos dados de treinamento e em um subconjunto aleatório de atributos.

Para os problemas de classificação, as decisões funcionam como um sistema de votação que combina as previsões de todas as árvores, definindo a previsão final pela classe mais frequente nessas previsões. No caso dos problemas de regressão, é utilizada a média para decidir a previsão final.

São utilizados hiperparâmetros para controlar o processo de treinamento como o número de árvores, que afeta a precisão e o tempo de treinamento, a profundidade máxima da árvore, que controla a complexidade modelo, e o número mínimo de amostras, que auxilia no controle do *overfitting*.

O algoritmo de *Random Forest*, ao utilizar diversas árvores de decisão treinadas com subconjuntos diferentes de dados, auxilia na redução do *overfitting* em comparação com o uso de uma única árvore. Dessa forma, também auxilia na identificação das variáveis mais relevantes e que mais contribuem na criação das árvores para representar a relação dos dados.

3 TRABALHOS RELACIONADOS

Existe uma ampla literatura de pesquisa sobre a dengue abordando estudos sobre a identificação de regiões críticas, detecção de surtos epidêmicos, predição de casos e correlação com variáveis climáticas e socioeconômicas (ALEIXO et al., 2022). Nesse capítulo está apresentada uma breve revisão dos trabalhos relacionados a predição de casos de dengue e da explicabilidade dos modelos propostos.

Segundo o artigo de revisão (SIRIYASATIEN et al., 2018), diversos desses trabalhos buscam desenvolver modelos preditivos para entender os fatores que influenciam a disseminação da doença e prever epidemias, utilizando análise estatística, matemática e aprendizado de máquina. Foram indicados 996 trabalhos com modelos para analisar epidemias de dengue segundo a base de dados SCOPUS. O levantamento feito nessa pesquisa mostrou que a técnica mais amplamente adotada para predição de dengue é a regressão, com 545 trabalhos no período de 2010 a 2017, seguida das séries temporais com 220 representantes no mesmo período.

Ainda conforme descrito no artigo, isso ocorre pois essas duas abordagens são mais simples de compreender e adequadas para, por exemplo, prever a saída no número de casos de dengue. Entretanto, as redes neurais, que apresenta a terceira técnica mais comum com 76 trabalhos, são mais complexas e dependem do método específico para avaliação. Seguidamente de técnicas mais simples, como as árvores de decisão, que contém 50 artigos.

Uma revisão apresentada no artigo (ROSTER; RODRIGUES, 2021) sobre a predição de dengue utilizando redes neurais apresenta a análise de 19 trabalhos, que em sua maioria utilizam dados históricos de incidência de dengue e características meteorológicas para realizar a predição com arquiteturas simples. O artigo destaca o potencial do uso de redes neurais mais complexas, como as RNNs e as CNNs, que são relativamente pouco exploradas, mas que demonstram promessas para pesquisas futuras. No entanto, não aborda diretamente o tema da explicabilidade na utilização desses modelos.

Existe uma falta na exploração da área de pesquisa de explicabilidade (XAI) nos trabalhos já existentes na predição de casos de dengue, principalmente quando utilizados modelos com saídas mais complexas, como redes neurais, como aponta o artigo (SIRIYASATIEN et al., 2018). Nesse contexto, os recentes trabalhos de (ALEIXO et al., 2022) e (PROME et al., 2024) buscam utilizar técnicas de aprendizado de máquina na explicação de seus modelos de predição e são detalhados abaixo.

O artigo (ALEIXO et al., 2022) apresenta um modelo explicável para prever o número de ocorrências de dengue na cidade do Rio de Janeiro, aplicando um algoritmo de método de regressão com árvores aumentadas (CatBoost) em comparação com o modelo autorregressivo integrado de médias móveis sazonal (SARIMA). As bases de dados

utilizadas para o treinamento do modelo foram do Sistema de Informação de Agravos de Notificação (SINAN) e do Instituto Brasileiro de Geografia e Estatística (IBGE), para dados meteorológicos e sociodemográficos. O período relativo aos dados de treinamento do modelo corresponde aos anos de 2016 a 2020, com uma janela de predição de 1 a 3 meses. A explicabilidade do modelo é dada através da técnica de SHAP (*SHapley Additive exPlanations*), que atribui a importância de cada variável de entrada para a predição do modelo.

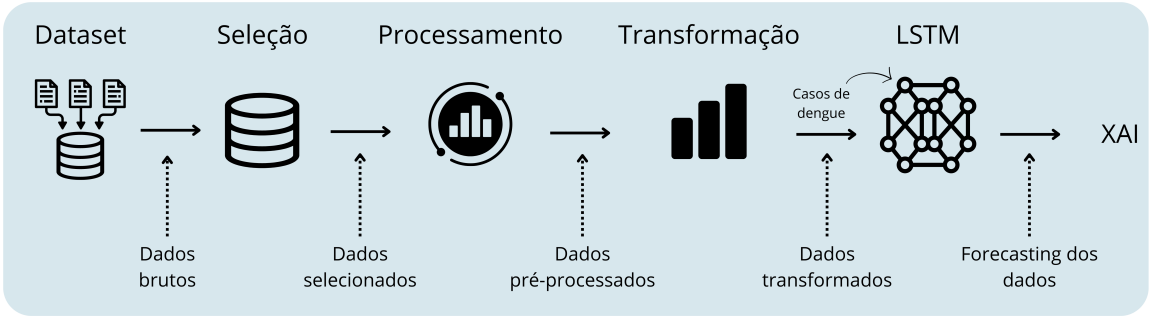
O artigo (PROME et al., 2024) buscou desenvolver um modelo para prever surtos de dengue nas regiões de Bangladesh, foram aplicados diversas técnicas de aprendizado de máquina como o modelo de regressão de vetores de suporte, que obteve os melhores resultados. O período utilizado para para análise parte de Janeiro de 2011 a Julho de 2021, com a base de dados do Serviços Meteorológicos e de Saúde de Bangladesh. A aplicação *Shapash* foi utilizada para realizar a explicabilidade do modelo, que também usa a técnica de SHAP.

Dessa forma, esse estudo busca também explorar a explicabilidade propondo um modelo para resolver problemas de XAI na predição de casos de dengue em modelos LSTM, através do uso de diferentes técnicas de aprendizado e classificação, em conjunto com as variáveis climáticas.

4 METODOLOGIA

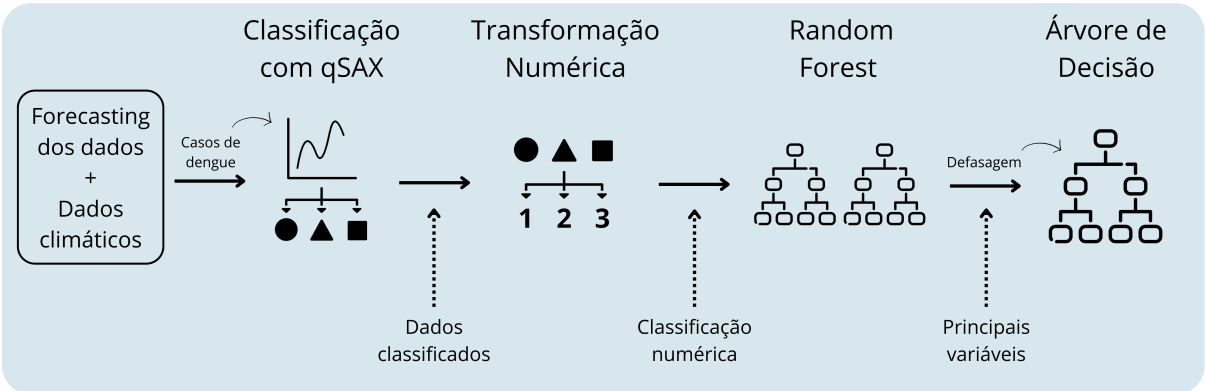
Esse capítulo apresenta a metodologia utilizada nesse trabalho, demonstrando as etapas desenvolvidas para alcançar os objetivos da pesquisa. São descritos os processos de coleta, seleção, pré-processamento e transformação do conjunto de dados, seguidamente da aplicação dos modelos de LSTM e SAX, e montagem das árvores. A Figura 3 mostra o processo de preparação dos dados e do *forecasting* dos casos de dengue pelo modelo LSTM. A Figura 4 apresenta o método criado para resolver o problema de XAI, utilizando técnicas de SAX e apresentando os resultados de forma explicável com árvores de decisão.

Figura 3 – Método de *Forecasting* dos casos de dengue.



Fonte: Própria autora

Figura 4 – Método proposto de Explanable AI (XAI).



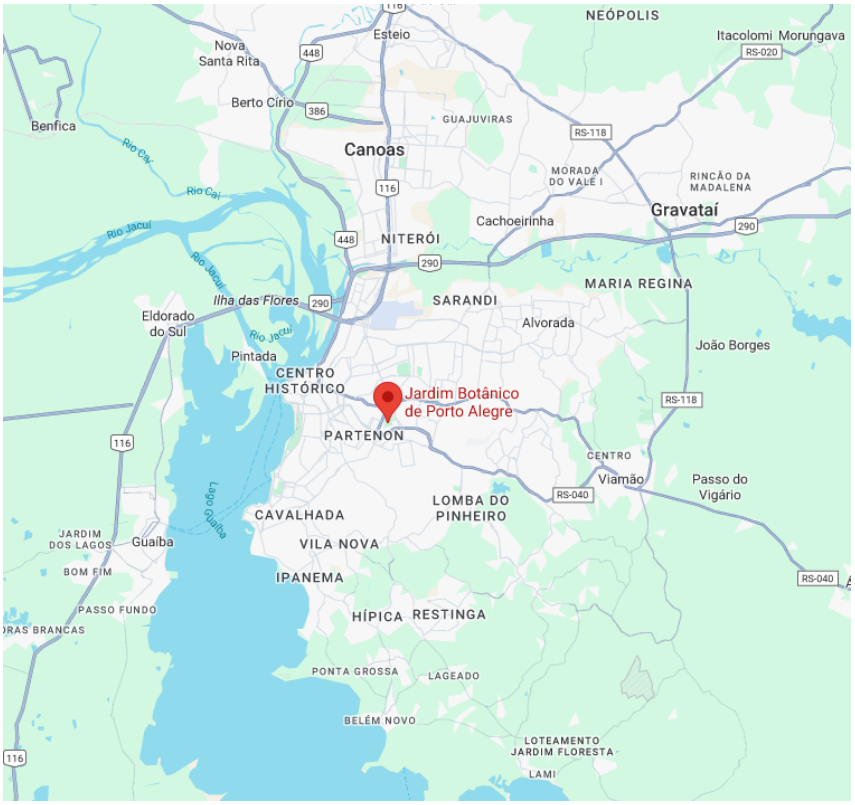
Fonte: Própria autora

4.1 CONJUNTO DE DADOS

A primeira etapa consistiu na escolha de uma região para análise dos casos de dengue no Brasil, dessa forma, foi escolhida a cidade de Porto Alegre devido a grande quantidade de registros nos últimos anos. A coleta de dados foi realizada através da plataforma InfoDengue[1], que disponibiliza uma API com os dados da situação epidemiológica da dengue de diversas cidades do país. Também foi utilizada em conjunto a API disponibilizada pelo Instituto Nacional de Meteorologia (INMET)[2] para obter variáveis climáticas como temperatura, umidade e precipitação, que são fatores conhecidos por influenciar a propagação do vírus, de acordo com o período selecionado.

Os dados coletados pela API do INMET referem-se à estação meteorológica automática Jardim Botânico (A801), localizada na região central de Porto Alegre, no bairro Jardim Botânico, apresentado na Figura 5. Esta é uma das duas estações na cidade que registram variáveis climáticas a cada hora, diferentemente das estações convencionais, que registram apenas três vezes ao dia. Vale ressaltar que os resultados são baseados apenas nos dados dessa estação, o que significa que podem não representar completamente as condições climáticas em outras partes da cidade.

Figura 5 – Localização da estação Jardim Botânico (A801).



Fonte: Google Maps. "Jardim Botânico de Porto Alegre". Captura de tela, 2024.

[1] Disponível em: https://info.dengue.mat.br/services/api
[2] Disponível em: https://portal.inmet.gov.br/

O conjunto de dados foi então unificado em um formato adequado para as próximas etapas em um arquivo CSV, com a realização da verificação da integridade dos registros e a correção de eventuais inconsistências nos dados referentes as semanas epidemiológicas da dengue, utilizando outras fontes de dados divulgados pela Prefeitura Municipal de Porto Alegre[3]. Após essa etapa, as variáveis foram revisadas, garantindo que todas as necessárias estivessem presentes e corretamente alinhadas temporalmente para as transformações realizadas posteriormente.

4.1.1 Seleção dos dados

Na etapa de seleção dos dados, foi escolhido o período de 2010 a 2023 que apresentava uma quantidade dados mais completa e é marcado pelo aumento do número de casos de dengue nos últimos anos. A capacidade da LSTM de resolver o problema de esquecimento temporal significa que ela pode aprender com sequências longas de dados, mantendo a memória de eventos passados relevantes para a previsão futura. Esse intervalo permite que os períodos anteriores melhorem o treinam do modelo e obtém melhor captura dos padrões encontrados, possibilitando melhores configurações para evitar o *overfitting*.

Continuamente, os dados referentes às condições climáticas foram escolhidos para relacionar com o número de casos de dengue, sendo eles a temperatura, umidade, ponto de orvalho, pressão, vento e precipitação, apresentados na Tabela 1. Essa escolha se deve à influência que esses fatores podem ter na propagação do vírus da dengue e na reprodução do mosquito vetor. Ao considerar esses dados, busca-se identificar padrões e correlações que possam ajudar a explicar e entender melhor a predição encontrada através do *forecasting* gerado pela LSTM.

Coluna	**Descrição**
Casos (N)	Número de casos de dengue registrados.
Data	Data do registro.
Temp. Ins. (C)	Temperatura instantânea em graus Celsius.
Umi. Ins. (%)	Umidade relativa instantânea em porcentagem.
Pto Orvalho Ins. (C)	Ponto de orvalho instantâneo em graus Celsius.
Pressao Ins. (hPa)	Pressão atmosférica instantânea em hectopascais.
Vel. Vento (m/s)	Velocidade do vento em metros por segundo.
Dir. Vento (m/s)	Direção do vento em metros por segundo.
Raj. Vento (m/s)	Rajada de vento em metros por segundo.
Chuva (mm)	Quantidade de chuva em milímetros.

Tabela 1 – Descrição das colunas do *dataset*.

[3]Disponível em: https://prefeitura.poa.br/sms/onde-esta-o-aedes/dados-de-porto-alegre

4.1.2 Pré-processamento dos dados

No pré-processamento, primeiramente foi realizada a filtragem das colunas que seriam utilizadas durante os processos, escolhidas na seleção dos dados, e mantendo os dados das semanas epidemiológicas que eram referentes ao número de casos de dengue na região. Ademais, foi realizada a limpeza dos dados relativos ao tratamento de valores ausentes, importante para garantir que o modelo não será influenciado por esses valores.

Por fim dessa etapa, foi escolhido realizar a análise de maneira mensal, para melhor compreensão do comportamento dos dados no decorrer dos anos. Para isso, foram somados os números de casos semanais das semanas epidemiológicas para obter o valor de casos mensais, e foi calculada a média dos valores das condições climáticas semanais. Alinhando novamente os dados temporalmente, obtemos um *dataset* com o número de casos de dengue mensais e suas médias das condições climáticas.

4.2 TRANSFORMAÇÃO DOS DADOS

A etapa de transformação dos dados consiste em preparar a entrada que será utilizada na rede neural que realiza o *forecast*, nesse caso a LSTM, sendo a coluna referente ao número de casos de dengue. A técnica utilizada é chamada de normalização ou padronização dos dados, realizada em dois passos descritos abaixo:

1. **Cálculo dos fatores de escala:** O primeiro passo é calcular os fatores de escala, que consistem na média (μ), calculada pela fórmula 4.1, e no desvio padrão (σ), apresentado na fórmula 4.2, da variável *Casos*.

$$\mu = \frac{1}{n}\sum_{i=1}^{n} x_i \tag{4.1}$$

$$\sigma = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(x_i - \mu)^2} \tag{4.2}$$

2. **Escalonamento dos dados:** Em seguida, os dados são escalonados utilizando a técnica de normalização. Cada valor de *Casos* é substituído por um valor escalonado, calculado pela fórmula 4.3, onde μ é a média e σ é o desvio padrão.

$$treinamento\_escalado = \frac{Casos - \mu}{\sigma} \tag{4.3}$$

Essa etapa é essencial, pois uniformiza a escala das variáveis, melhora a eficiência e precisão dos modelos de aprendizado de máquina, facilitando a identificação de padrões.

Assim, reduz o impacto de *outliers*, ajusta os dados aos requisitos específicos dos algoritmos e prepara adequadamente para a modelagem, contribuindo significativamente para a eficácia dos resultados preditivos do modelo.

4.3 MODELOS E MINERAÇÃO

4.3.1 Aplicação da LSTM

O primeiro passo para a aplicação da rede neural LSTM consiste em preparar os dados do número de casos de dengue, já escalonados na etapa de transformação, para treinar o modelo. Para isso, é preciso definir qual será o período de predição realizado pelo modelo, que também define quantos valores anteriores a rede deve se basear para cada predição. Assim, para fazer o *forecasting* de x predições, cada uma se baseia em uma janelas de x pontos de dados. Abaixo está um exemplo simplificado de como é realizada a montagem do conjunto de treino:

- Vamos supor que a série temporal tenha 15 pontos de dados: $[1, 2, 3, \ldots, 15]$ e queremos prever os próximos 6 meses.
- Precisamos criar janelas de 6 pontos de dados para prever os próximos 6 valores. O conjunto de treinamento ficará assim:

$$
\begin{aligned}
[1, 2, 3, 4, 5, 6] &\rightarrow [7, 8, 9, 10, 11, 12] \\
[2, 3, 4, 5, 6, 7] &\rightarrow [8, 9, 10, 11, 12, 13] \\
[3, 4, 5, 6, 7, 8] &\rightarrow [9, 10, 11, 12, 13, 14] \\
[4, 5, 6, 7, 8, 9] &\rightarrow [10, 11, 12, 13, 14, 15]
\end{aligned}
$$

- Os preditores (X) e o alvos (Y) são iguais a:

$$
X = \begin{bmatrix} 1 & 2 & 3 & 4 & 5 & 6 \\ 2 & 3 & 4 & 5 & 6 & 7 \\ 3 & 4 & 5 & 6 & 7 & 8 \\ 4 & 5 & 6 & 7 & 8 & 9 \end{bmatrix} \quad Y = \begin{bmatrix} 7 & 8 & 9 & 10 & 11 & 12 \\ 8 & 9 & 10 & 11 & 12 & 13 \\ 9 & 10 & 11 & 12 & 13 & 14 \\ 10 & 11 & 12 & 13 & 14 & 15 \end{bmatrix}
$$

Dessa forma, como nesse trabalho queremos fazer o *forecasting* do ano de 2024, preditores e janelas são definidos para 12 meses e o conjunto de treinamento do modelo é montado como demonstrado no exemplo.

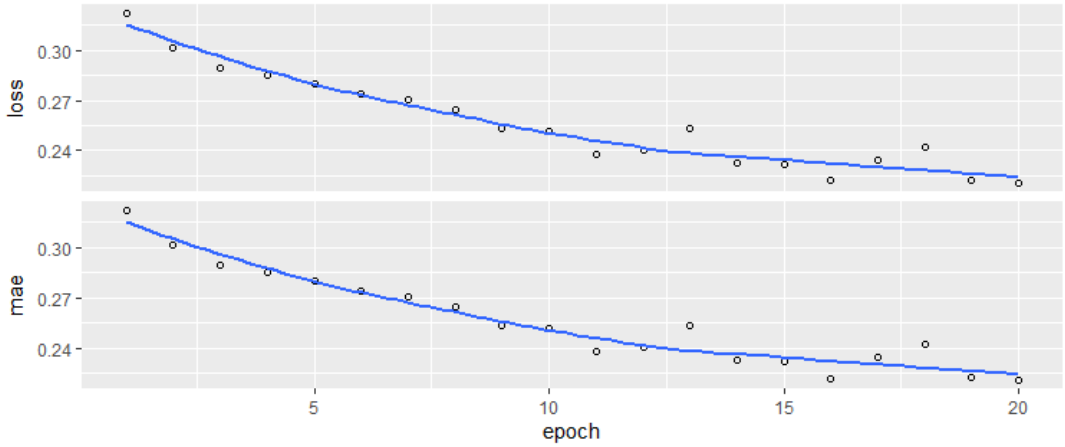
Na sequência, com o conjunto de treinamento formado, é feita a transformação das matrizes obtidas em vetores de 3 dimensões, formato correto para a entrada da rede neural. Também é definido o conjunto de teste, correspondente as últimas 12 observações do número de casos de dengue do *dataset*, transformada igualmente em um vetor de 3 dimensões.

O modelo da rede neural LSTM foi desenvolvido utilizando as bibliotecas *Keras* e *Tensorflow* da linguagem de programação R, no ambiente RStudio. A implementação completa, juntamente com suas referências, está disponível na nota de rodapé[4]. Foram utilizadas duas camadas LSTM para processar as sequências de dados e aprender as dependências temporais, com número de neurônios igual a 50, duas camadas de *dropout*, onde 50% dos neurônios serão desativados aleatoriamente durante cada passo de treinamento, e uma camada densa para mapear as saídas recorrentes.

As camadas e parâmetros utilizados foram ajustados para evitar o *overfitting* do modelo, para isso, foi utilizado como medida de perda o erro médio absoluto, calculado pela Fórmula 4.4, onde n é o número de exemplos, $ytrue$ são os valores reais e $ypred$ são os valores previstos pelo modelo. O MAE (Mean Absolute Error) indica o quão próximas as previsões do modelo estão dos valores reais, sendo uma métrica de erro absoluto. A Figura 6 demonstra o histórico do modelo treinado, com os valores de MAE.

$$MAE = \frac{1}{n}\sum_{i=1}^{n} |y_{\text{true}} - y_{\text{pred}}| \tag{4.4}$$

Figura 6 – Histórico de treinamento do modelo LSTM.

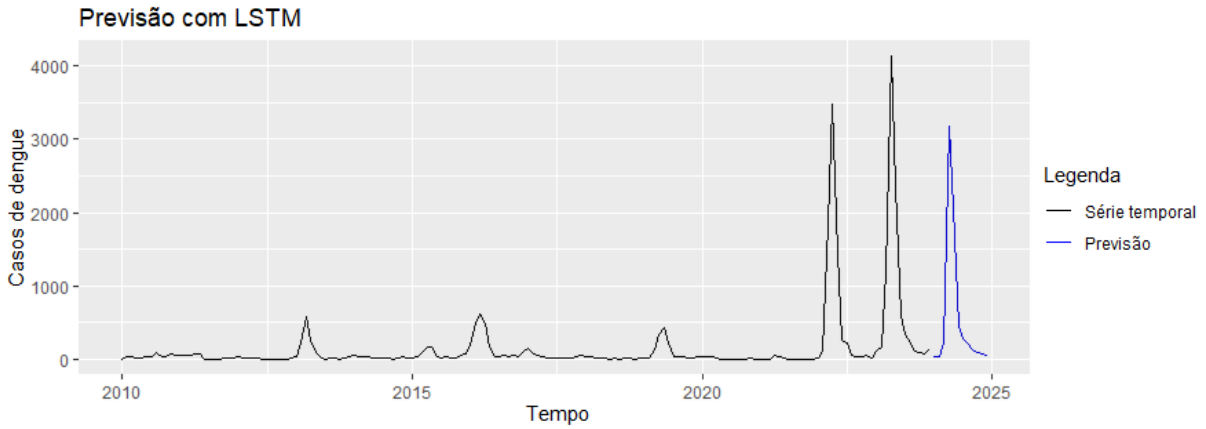


Fonte: Própria autora

Enfim, obtemos o *forecasting* do modelo LSTM referente aos 12 meses seguintes ao final do conjunto para o número de casos de dengue em Porto Alegre, representado na Figura 7, onde são apresentados no eixo x os anos, no eixo y o número de casos e destacado em azul a predição dos dados.

[4]https://github.com/AndrizaCampanhol/AndrizaCampanhol-Explainable-Model-for-Dengue-Prediction

Figura 7 – Forecasting do modelo LSTM.



Fonte: Própria autora

4.3.2 Aplicação das técnicas de SAX

Para realizar a classificação dos dados obtidos pelo *forecasting* e demais registros do *dataset*, foi utilizada a função de *Discretize* na plataforma *Symbols!*[5], onde foram aplicadas as técnicas de SAX e qSAX nos dados, selecionando duas classes. Essas técnicas de classificação foram apresentadas na Seção 2.3 do Capítulo 2.

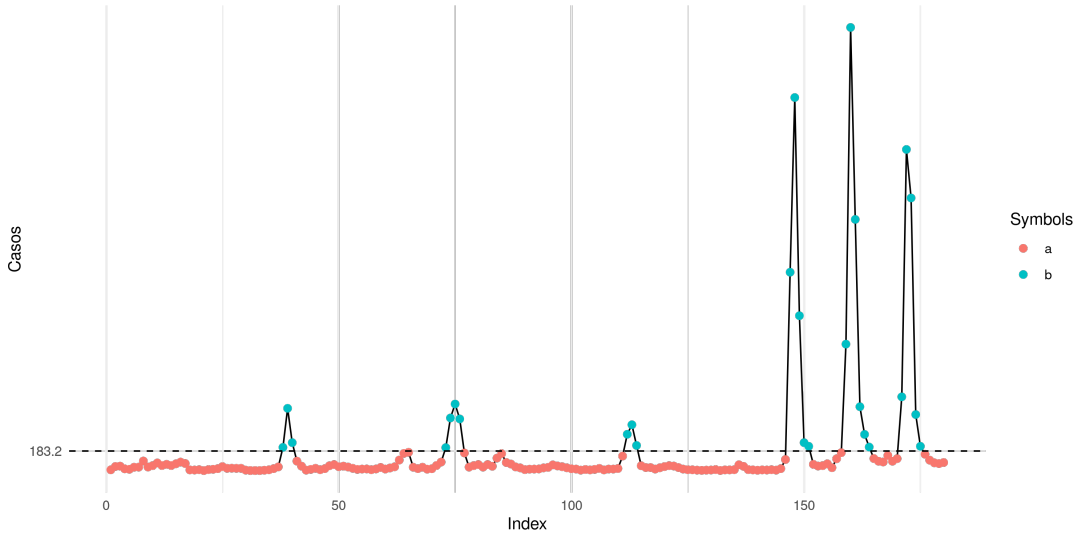
A abordagem inicial deste trabalho utiliza dois casos, com duas e cinco classes, para aplicar a classificação com SAX e qSAX. A escolha inicial de duas classes buscou simplificar a demonstração da técnica para a detecção e explicação de “explosões“ de casos de dengue. No entanto, o uso de cinco classes pode proporcionar uma análise mais precisa, capaz de identificar variações mais sutis nos padrões de casos ao longo do tempo. É importante ressaltar que essas técnicas são altamente flexíveis e permitem a utilização de um número variado de classes ou símbolos.

As Figuras 8 e 9 apresentam a classificação obtida utilizando as técnicas de SAX e qSAX, respectivamente, onde a representa baixo número de casos de dengue e b representa maior número de casos de dengue. Ao aplicar a primeira técnica, obtemos uma melhor classificação dos picos de casos que ocorrem sazonalmente, porém, como o SAX leva em consideração que a distribuição é normalizada, a grande maioria dos casos está classificado próximo de 0. Entretanto, empregando a técnica de qSAX, obtemos uma distribuição balanceada dos anos que apresentam casos de dengue expressivos, independente do número de símbolos, sendo então a técnica escolhida.

A Figura 10 mostra o resultado para 5 classes com SAX, enquanto a Figura 11 com qSAX. Nota-se que os picos de casos de dengue estão melhor identificados em comparação com a classificação utilizando apenas dois símbolos e que o SAX não consegue identificar todas as 5 classes, reforçando a escolha do qSAX.

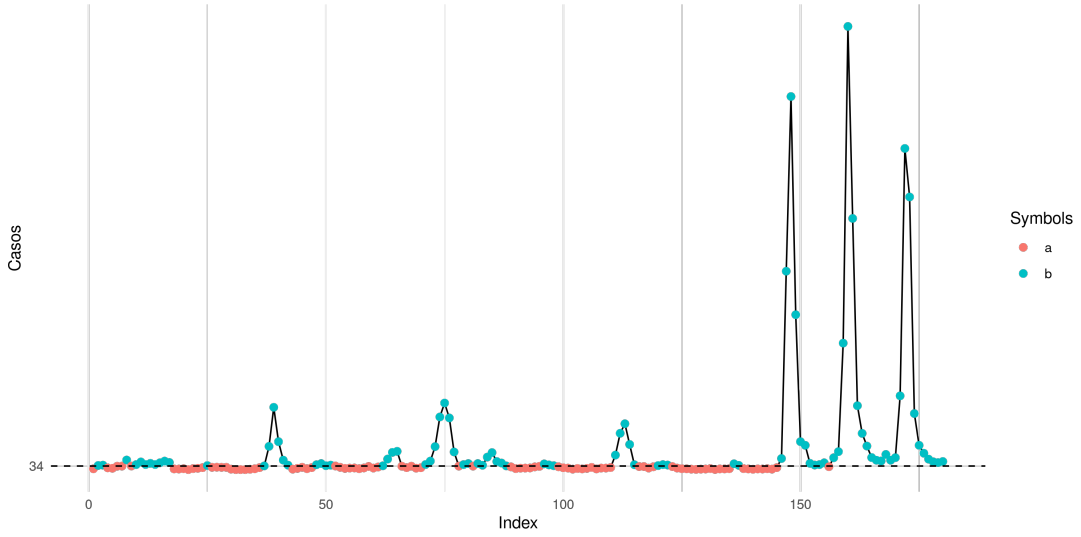
[5]Disponível em: https://esilveira.shinyapps.io/symbols/

Figura 8 – Discretização realizada com SAX em 2 classes.



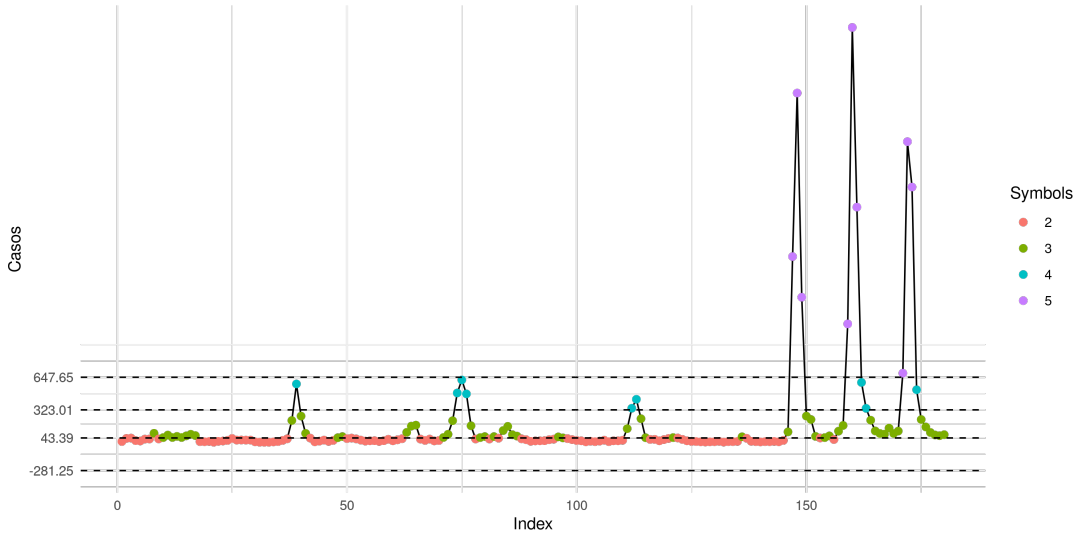
Fonte: Própria autora.

Figura 9 – Discretização realizada com qSAX em 2 classes.



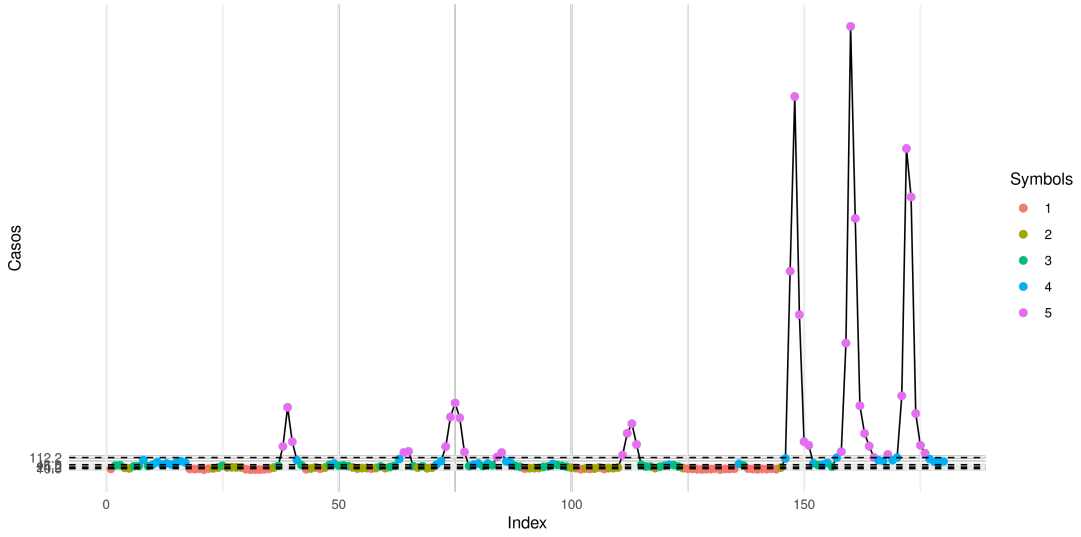
Fonte: Própria autora.

Figura 10 – Discretização realizada com SAX em 5 classes.



Fonte: Própria autora.

Figura 11 – Discretização realizada com qSAX em 5 classes.



Fonte: Própria autora.

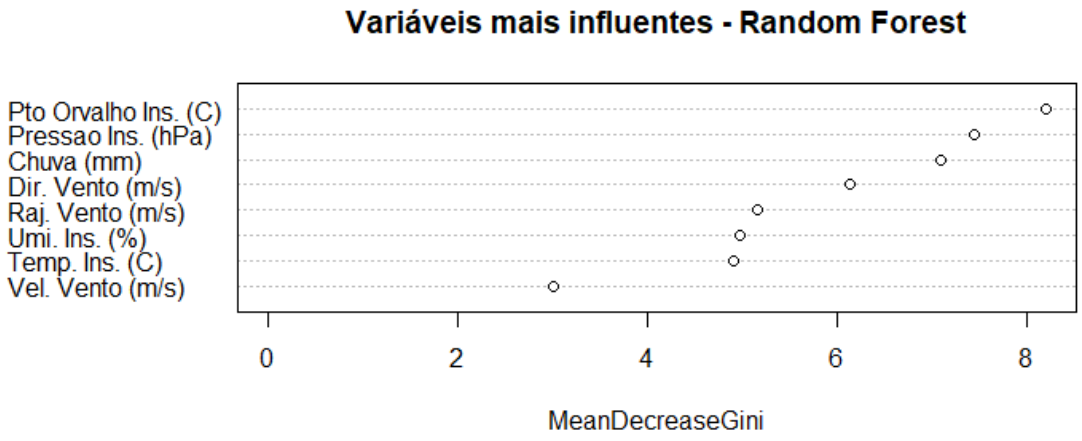
4.3.3 Principais variáveis com *Random Forest*

Com os dados classificados pelo qSAX, que obtiveram exito em identificar os picos de casos, primeiramente foi realizada a transformação numérica, passando os símbolos para números, que são melhor reconhecidos pelos modelos computacionais. Assim, foi aplicado o algoritmo de *Random Forest* para encontrar as variáveis mais influentes para a classificação dos dados em cada classe, apresentado na Figura 12, calculado através da Redução Média de Gini.

O Gini é utilizado para avaliar a importância de cada variável no processo de tomada de decisão do modelo, ele indica o quanto cada variável contribui para a pureza das folhas da árvore de decisão, ou seja, o quanto ela contribui para a capacidade do modelo de fazer previsões precisas.

Para a configuração do algoritmo, foram utilizadas 500 árvores, com 8 características por árvore, sendo elas: Temperatura, Umidade, Ponto de Orvalho, Pressão, Velocidade do Vento, Direção do Vento, Rajadas de Vento e Chuva. Os conjuntos de treino e teste foram divididos com 80% e 20% dos dados, respectivamente, correspondendo ao período de aprendizagem da rede neural.

Figura 12 – Principais variáveis com Random Forest.



Fonte: Própria autora

4.3.4 Visualização em Árvore de Decisão

Na sequência, foram utilizadas as três variáveis mais influentes na classificação como entrada para o algoritmo de *Decision Tree*, utilizando o método *"class"*. Dessa forma, podemos visualizar a relação das classes encontradas com as variáveis climáticas referentes aos meses observados, refletindo ao período de treino do modelo LSTM.

Foram testadas diferentes configurações de árvores para encontrar a de melhor acurácia no caso proposto. Inicialmente, foram montadas considerando apenas os valores mais influentes. Com base nesses resultados, os valores das colunas foram atrasados temporalmente com intuito de obter um modelo que incorporasse o comportamento das variáveis nos meses anteriores, para explorar suas dependências temporais e buscar obter resultados mais precisos.

As novas colunas foram criadas utilizando o processo de “defasagem“ dos dados em dois meses, para as três variáveis mais influentes encontradas anteriormente: Ponto de Orvalho, Pressão e Chuva. A Tabela 2 demonstra os valores defasados da variável de chuva como exemplo, onde “Chuva (mm)-1“ e “Chuva (mm)-2“ representam o primeiro e o segundo mês anterior, respectivamente.

Data (mês)	**Chuva (mm)**	**Chuva (mm)-1**	**Chuva (mm)-2**
2023-08-31	77.0	210.8	247.6
2023-09-30	394.8	77.0	210.8
2023-10-31	110.0	394.8	77.0
2023-11-30	342.8	110.0	394.8
2023-12-31	88.6	342.8	110.0

Tabela 2 – Defasagem da variável Chuva.

5 RESULTADOS

Neste capítulo, são apresentados as análises para a construção das árvores utilizando 2 e 5 classes, comparando os resultados com e sem a aplicação da técnica de defasagem. As porcentagens relativas aos nós da árvore representam a distribuição dos dados de treino para a classe alvo daquele nó, ou seja, o total de observações que chegou naquele ponto da árvore, enquanto os valores entre 0 e 1 representam a perda esperada pelo treinamento em cada nó.

5.1 ÁRVORES DE DUAS CLASSES

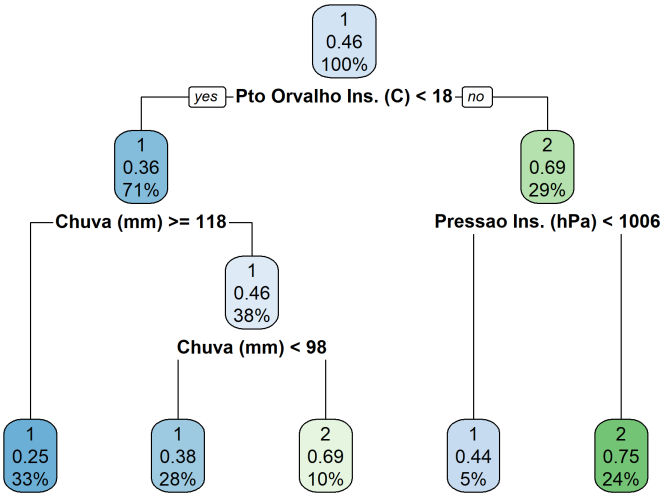
As árvores de duas classes nesta subseção classificam nós como 1 para baixo número de casos de dengue e nós como 2 para maior número de casos de dengue.

A Figura 13 demonstra a árvore de decisão com duas classes e profundidade 3, sem utilizar a técnica de defasagem. A Tabela 3 demonstra a matriz de confusão dessa árvore, que obteve uma acurácia de 0.6964, indicando que 69.64% das previsões foram corretas ao realizar previsões de classificação nessa árvore.

Predição \ Referência	**Classe 1**	**Classe 2**
Classe 1	76	36
Classe 2	15	41

Tabela 3 – Matriz de confusão da árvore com duas classes e profundidade 3.

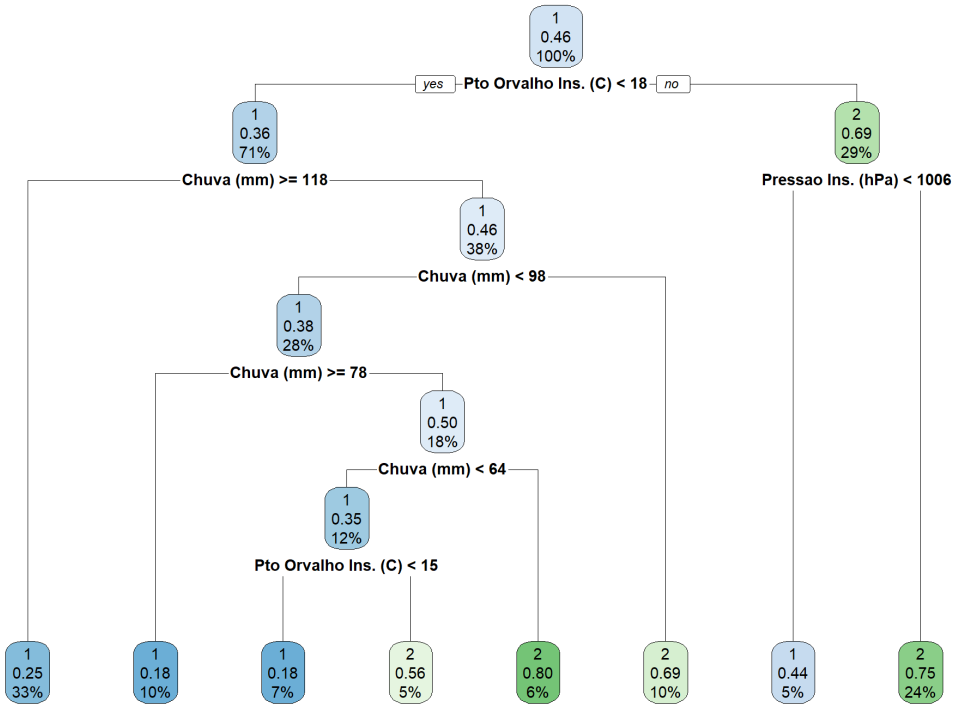
Figura 13 – Árvore de decisão com duas classes e profundidade 3.



Fonte: Própria autora.

A Figura 14 demonstra a árvore de decisão com duas classes e profundidade 6, sem utilizar a técnica de defasagem. A Tabela 4 demonstra a matriz de confusão dessa árvore, que obteve uma acurácia de 73.81%.

Figura 14 – Árvore de decisão com duas classes e profundidade 6.



Fonte: Própria autora.

Predição \ Referência	Classe 1	Classe 2
Classe 1	70	23
Classe 2	21	54

Tabela 4 – Matriz de confusão da árvore com duas classes e profundidade 6.

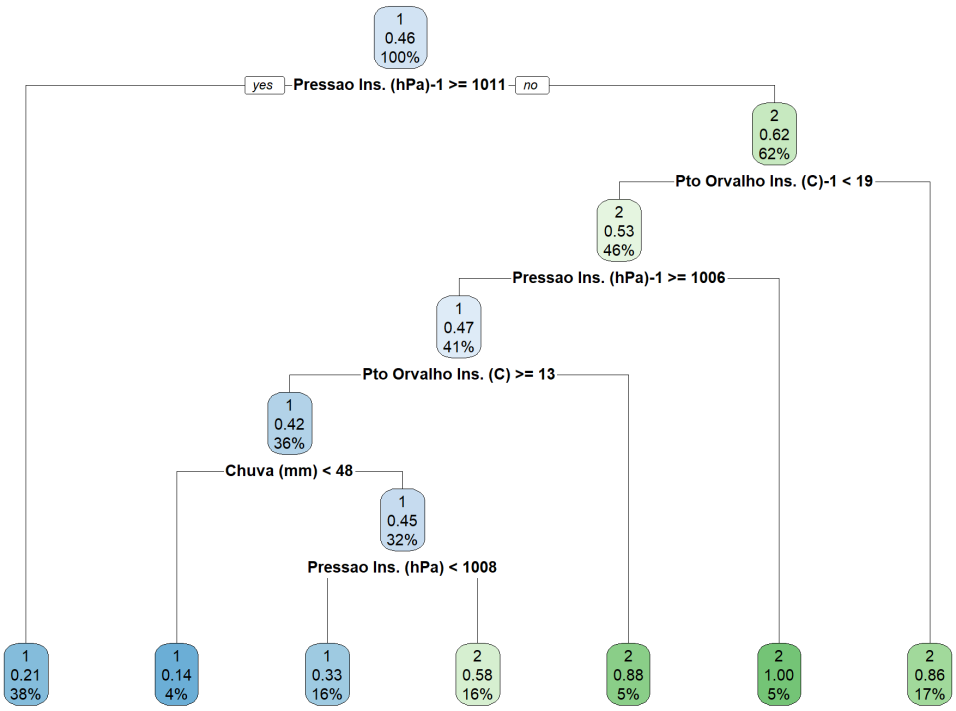
Na sequência, com objetivo de ampliar a análise da influência das variáveis encontradas, foi utilizada a técnica de defasagem dos dados em um e dois meses. Dessa forma, sendo possível utilizar também valores referentes aos meses anteriores na montagem.

A Figura 15 demonstra a árvore de duas classes, profundidade 6 e defasagem de um mês. A Tabela 5 demonstra a matriz de confusão da árvore, com acurácia de 76.65%.

Predição \ Referência	Classe 1	Classe 2
Classe 1	74	23
Classe 2	16	54

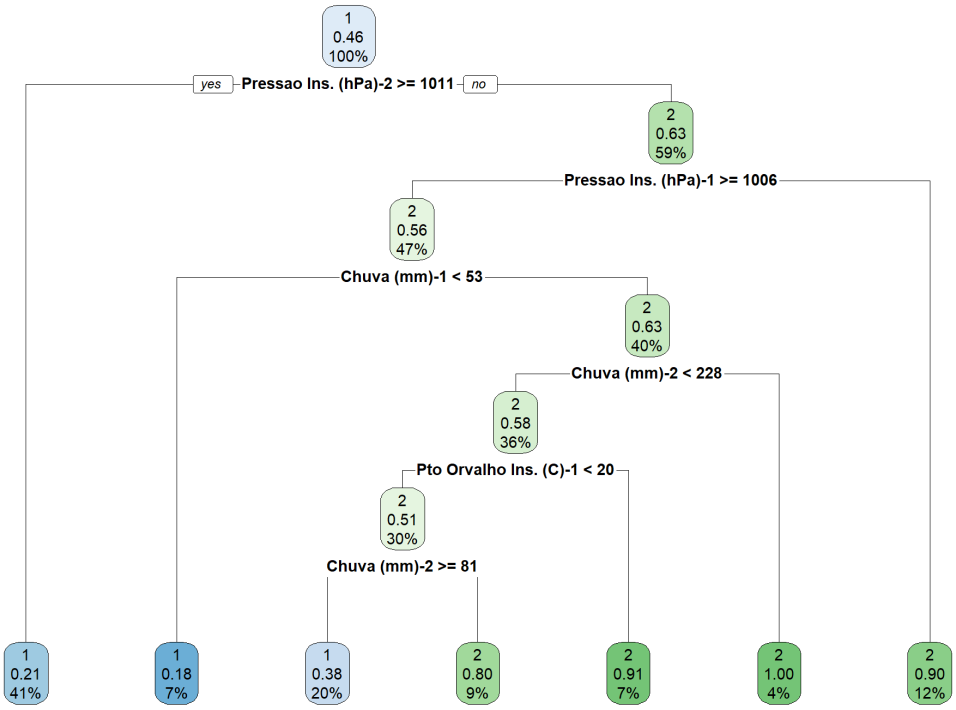
Tabela 5 – Matriz de confusão da árvore com duas classes e profundidade 6.

Figura 15 – Árvore de decisão com duas classes, profundidade 6 e defasagem 1.



Fonte: Própria autora.

Figura 16 – Árvore de decisão com duas classes, profundidade 6 e defasagem 2.



Fonte: Própria autora.

A Figura 16 demonstra a árvore de duas classes, profundidade 6 e defasagem de dois meses. A Tabela 6 demonstra a matriz de confusão da árvore, de acurácia 78.92%.

Predição \ Referência	**Classe 1**	**Classe 2**
Classe 1	84	29
Classe 2	6	47

Tabela 6 – Matriz de confusão da árvore com duas classes e profundidade 6.

5.2 ÁRVORES DE CINCO CLASSES

As árvores de cinco classes nesta subseção classificam nós de 1 até 5, de baixo para alto número de casos de dengue, respectivamente, também sendo identificados na utilização de diferentes cores na imagem. Como a defasagem demonstrou melhora nas métricas quando utilizando 2 classes, as árvores de 5 classes já são montadas utilizando a defasagem de 2 meses.

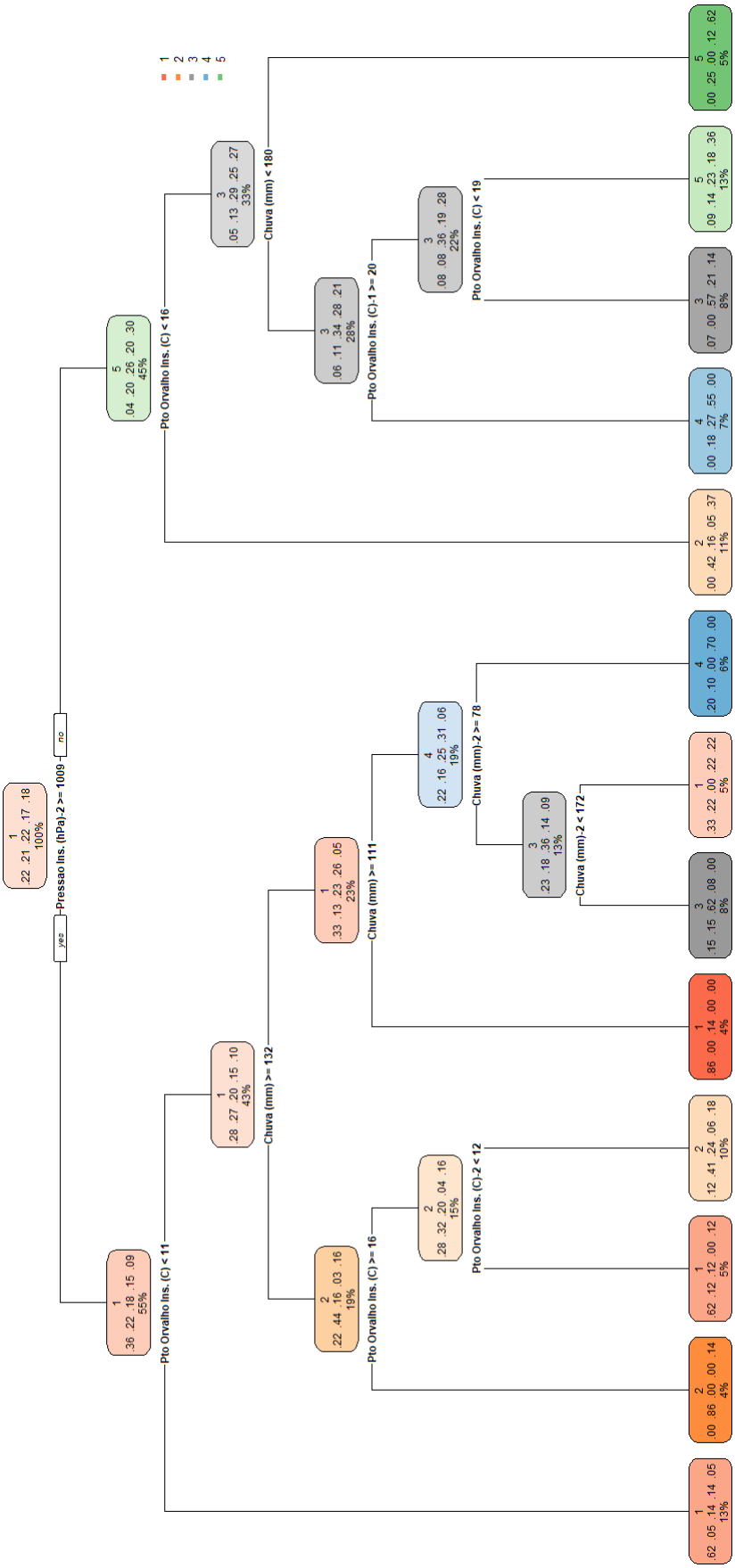
Predição \ Referência	**Classe 1**	**Classe 2**	**Classe 3**	**Classe 4**	**Classe 5**
Classe 1	27	4	5	5	4
Classe 2	2	21	7	2	11
Classe 3	3	2	16	4	2
Classe 4	2	3	3	13	0
Classe 5	2	5	5	5	13

Tabela 7 – Matriz de confusão da árvore com 5 classes.

A Figura 17 demonstra a árvore de decisão obtida nessa etapa ao utilizar 5 classes na classificação. Nesse caso, a árvore se torna complexa para realizar sua análise e interpretação, com uma acurácia baixa de 54.22%, em relação a matriz de confusão apresentada na Tabela 7.

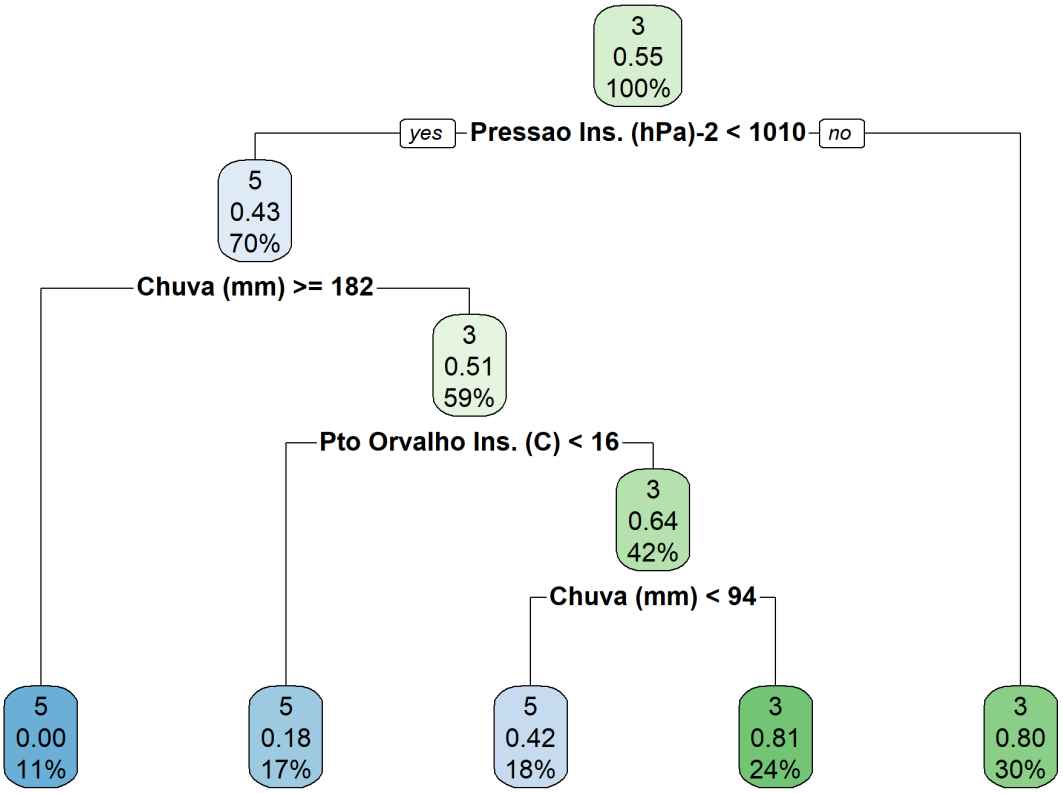
Como o objetivo desse trabalho é identificar os picos nos casos de dengue, podemos utilizar a simplificação da árvore através da diminuição da análise dos fatores em sua montagem, dessa forma, obtemos a árvore da Figura 18, utilizando as classes 3 e 5. Nessa árvore podemos identificar melhor as variáveis para a classificação do grande número de casos, representado pela classe 5, e que obter uma acurácia de 78.79%, na matiz de confusão apresentada na Tabela 8.

Figura 17 – Árvore de decisão com 5 classes.



Fonte: Própria autora.

Figura 18 – Árvore de decisão com 5 classes e simplificação.



Fonte: Própria autora.

Predição \ Referência	Classe 5	Classe 3
Classe 5	23	7
Classe 3	7	29

Tabela 8 – Matriz de confusão da árvore com 5 classes e simplificação.

5.3 DISCUSSÃO DOS RESULTADOS

Através da aplicação do modelo proposto nesse trabalho, primeiramente foi possível realizar a predição do número de casos de dengue através de uma rede neural LSTM no ano de 2024, que obteve resultados satisfatórios de treinamento e predição, analisados pelo parâmetro de MAE (Ver Subseção 4.3.1). Seguidamente da aplicação das técnicas de SAX e qSAX para discretização dos casos de dengue obtidos no *forecasting*, apresentando uma distribuição balanceada dos dados classificados (Ver Subseção 4.3.2), importante para as etapas seguintes de explicabilidade.

Dessa forma, foram criadas árvores baseadas nas variáveis mais influentes encontradas pelo algoritmo de *Random Forest*: Ponto de Orvalho, Pressão e Chuva, calculadas através da Redução Média de Gini (Ver Figura 12). Na sequência, com objetivo de ampliar a análise das dependências temporais das variáveis encontradas, foi utilizada a técnica de defasagem dos dados em dois meses. Assim, obtendo como resultado explicativo as árvores de decisão, que são discutidas abaixo.

As métricas para avaliação das árvores utilizadas são referentes à acurácia, que representa a proporção de previsões corretas em relação ao total de observações, e ao *F1-score*, que é a média harmônica entre dois parâmetros: precisão e sensibilidade. A precisão refere-se a proporção de resultados positivos identificados corretamente, e a sensibilidade refere-se aos positivos reais identificados corretamente. Nesse trabalho, os valores positivos correspondem a identificação da classe que representa os picos no número de casos de dengue.

A partir das métricas apresentados na Tabela 9, a árvore da Figura 16 foi a mais adequada para a explicabilidade da predição de casos de dengue para 2 classes, com 78.92% de acurácia e 75.44% de *F1-score*, e a árvore da Figura 18 para 5 classes, com 78.79% de acurácia e 76.67% de *F1-score*, utilizando o modelo proposto. Também é possível identificar que a utilização da técnica de defasagem auxiliou na melhor visualização das características mais influentes para a classificação desses dados temporalmente.

Além disso, as árvores com 5 classes apresentaram uma acurácia menor em relação as de 2 classes, porém, a métrica de *F1-Score* mostra que ao utilizar a simplificação, o modelo apresentou capacidade maior de classificar corretamente as explosões de casos, possivelmente causado pelo volume de dados. Dessa maneira, uma forma de obter resultados mais precisos pode vir na aplicação do método em problemas com maior número de registros temporais disponíveis.

Nº de Classes	**Profundidade**	**Defasagem**	**Acurácia (%)**	**F1-score (%)**
2	3	0	69.64	61.73
2	6	0	73.81	69.77
2	6	1	76.65	74.03
2	6	2	**78.92**	**75.44**
5	6	2	54.22	54.17
5	4	2	**78.79**	**76.67**

Tabela 9 – Resultados dos modelos de árvore de decisão.

Nesse contexto, os resultados iniciais são encorajadores para continuar explorando a utilização de técnicas de classificação na explicabilidade do modelo, visto que a literatura predominantemente utiliza de técnicas de regressão para abordar o problema (Ver Seção 3). Entretanto, visto que não há outra aplicação no mesmo conjunto de dados, com a mesma janela de predição ou utilização do conjunto de técnicas, não cabe realizar a comparação com outros modelos nesse momento.

A utilização de técnicas de classificação busca também melhorar a avaliação do modelo visto que um problema relatado na literatura (Ver Seção 3) é a dificuldade de avaliar metricamente modelos com dados relacionados aos casos de dengue devido às suas características sazonais, que acabam reduzindo a precisão dos resultados obtidos pelo comportamento temporal dos dados.

Ademais, as árvores com defasagem, Figuras 15, 16 e 18, demonstraram que os valores das três variáveis nos meses anteriores foram mais relevantes para a classificação dos casos, evidenciando que o comportamento climático nesse período pode causar um maior pico de casos de dengue nos meses seguintes.

De maneira geral, os resultados encontrados nessa pesquisa, utilizando as variáveis climáticas na explicabilidade da predição de casos de dengue, podem colaborar com as descobertas na área da saúde sobre a influência desses fatores no aumento da presença da doença e do mosquito, como apresentadas nos artigos (NAISH et al., 2014) e (ROCKLÖV; TOZAN, 2019).

6 CONCLUSÃO

Devido ao crescimento significativo da dengue nos últimos anos e a falta de exploração de modelos explicativos na predição da doença, esse trabalho buscou propor um modelo para resolver problemas de XAI na predição de casos de dengue na cidade de Porto Alegue. O modelo proposto utiliza dados temporais das condições climáticas para proporcionar uma compreensão mais clara dos fatores que contribuem para o aumento dos casos de dengue.

Na metodologia de aplicação do modelo proposto, a predição dos casos de dengue foi feita utilizando uma rede neural LSTM. Em seguida, os dados foram discretizados utilizando as técnicas de SAX e qSAX, a fim de detectar a explosão de casos na série temporal. Para realizar a explicabilidade, o algoritmo *Random Forest* foi utilizado para destacar as variáveis climáticas mais significativas na predição. Por fim, foram desenvolvidas árvores de decisão com diferentes configurações, e avaliadas as mais adequadas na explicação, sendo elas apresentadas nas Figuras 16 e 18 para 2 e 5 classes, respectivamente.

Os resultados das árvores de decisão, com a maior acurácia atingida de 78.92%, mostraram ser promissores para a contínua exploração da utilização de técnicas de classificação para modelos de explicabilidade, especialmente devido a identificação dos padrões sazonais apresentados pela doença, apesar da precisão dos resultados ter sido prejudicada pelo volume de dados disponível.

Por fim, os resultados relativos a análise do comportamento climático associado a predição de casos de dengue mostrou que o Ponto de Orvalho, a Pressão e a Chuva são relevantes na classificação dos picos de casos, podendo assim colaborar com as descobertas na área da saúde sobre a influência desses fatores climáticos na forte presença da doença e na proliferação do mosquito.

6.1 TRABALHOS FUTUROS

Para os trabalhos futuros, primeiramente, a possibilidade de aplicação do método proposto para explicabilidade da saída de modelos LSTM em diferentes conjuntos de dados relativos a dengue e as condições climáticas. Ademais, podendo abordando diversos problemas de explicabilidade no contexto de inteligência artificial e na predição de séries temporais.

Também seria interessante explorar o aprimoramento do modelo, utilizando, por exemplo, maiores níveis de discretização e maior volume de dados, buscando obter resultados mais precisos e maiores características do comportamento sazonal da dengue associada aos fatores climáticos.

Outra possibilidade de pesquisa seria adaptar o método proposto utilizando diferentes redes neurais para realizar a predição do número de casos de dengue em comparação com o uso da rede LSTM. Também podendo variar as técnicas de classificação, aplicadas com o SAX e qSAX no trabalho, para investigar essa abordagem menos utilizada na literatura.

Além disso, o modelo pode ser comparado com outras técnicas e modelos de explicabilidade em predição de casos de dengue, assim como em diferentes cenários em que pode ser aplicado, podendo proporcionar uma compreensão mais ampla de suas vantagens e limitações.

REFERÊNCIAS

AHMED, K. F. et al. An interpretable framework for predicting type 2 diabetes using ml and explainable ai. In: IEEE. **2023 26th International Conference on Computer and Information Technology (ICCIT)**. Cox's Bazar, Bangladesh, 2023. p. 1–6.

ALEIXO, R. et al. Predicting dengue outbreaks with explainable machine learning. In: IEEE. **2022 22nd IEEE International Symposium on Cluster, Cloud and Internet Computing (CCGrid)**. Taormina, Italy, 2022. p. 940–947.

BREIMAN, L. . Random forests. **Machine learning**, Springer, v. 45, p. 5–32, 2001.

BREIMAN, L. Bagging predictors. **Machine learning**, Springer, v. 24, p. 123–140, 1996.

BREIMAN, L. et al. **Classification and Regression Trees**. 1st. ed. New York: Chapman and Hall/CRC, 1984. 368 p. ISBN 9781315139470.

DOŠILOVIĆ, F. K.; BRČIĆ, M.; HLUPIĆ, N. Explainable artificial intelligence: A survey. In: IEEE. **2018 41st International convention on information and communication technology, electronics and microelectronics (MIPRO)**. Opatija, Croatia, 2018. p. 0210–0215.

GOVERNO FEDERAL. MINISTÉRIO DA SAÚDE. **Dengue - Ministério da Saúde**. 2024. https://www.gov.br/saude/pt-br/assuntos/saude-de-a-a-z/d/dengue. (Accessed on 04/16/2024).

GUIDOTTI, R. et al. Principles of explainable artificial intelligence. **Explainable AI Within the Digital Transformation and Cyber Physical Systems: XAI Methods and Applications**, Springer, p. 9–31, 2021.

HOCHREITER, S.; SCHMIDHUBER, J. Long short-term memory. **Neural computation**, MIT press, v. 9, n. 8, p. 1735–1780, 1997.

KLOSKA, M.; ROZINAJOVA, V. Distribution-wise symbolic aggregate approximation (dw-sax). In: SPRINGER. **Intelligent Data Engineering and Automated Learning–IDEAL 2020: 21st International Conference**. Guimaraes, Portugal, 2020. p. 304–315.

LIN, J. et al. Experiencing sax: a novel symbolic representation of time series. **Data Mining and knowledge discovery**, Springer, v. 15, p. 107–144, 2007.

NAISH, S. et al. Climate change and dengue: a critical and systematic review of quantitative modelling approaches. **BMC infectious diseases**, Springer, v. 14, p. 1–14, 2014.

PROME, S. S. et al. Prediction of dengue cases in bangladesh using explainable machine learning approach. In: IEEE. **2024 International Conference on Inventive Computation Technologies (ICICT)**. Lalitpur, Nepal, 2024. p. 1–5.

QUINLAN, J. R. Induction of decision trees. **Machine learning**, Springer, v. 1, p. 81–106, 1986.

ROCKLÖV, J.; TOZAN, Y. Climate change and the rising infectiousness of dengue. **Emerging Topics in Life Sciences**, Portland Press Ltd., v. 3, n. 2, p. 133–142, 2019.

ROSTER, K.; RODRIGUES, F. A. Neural networks for dengue prediction: a systematic review. **arXiv preprint arXiv:2106.12905**, 2021.

SECRETARIA MUNICIPAL DE SAÚDE DE PORTO ALEGRE. **Saúde confirma 1.810 casos de dengue na Capital em 2024 | Prefeitura de Porto Alegre**. 2024. https://prefeitura.poa.br/sms/noticias/saude-confirma-1810-casos-de-dengue-na-capital-em-2024. (Accessed on 04/16/2024).

SILVEIRA, E.; ASSUNÇÃO, J.; EMMENDORFER, L. Quantile symbolic aggregate approximation: A guaranteed equiprobable sax. In: SBC. **Anais do XXXVIII Simpósio Brasileiro de Bancos de Dados**. Belo Horizonte/MG, 2023. p. 396–401.

SIRIYASATIEN, P. et al. Dengue epidemics prediction: A survey of the state-of-the-art based on data science processes. **IEEE Access**, IEEE, v. 6, p. 53757–53795, 2018.

SOROUSH, K.; RAJI, M.; GHAVAMI, B. Compressing deep neural networks using explainable ai. In: IEEE. **2023 13th International Conference on Computer and Knowledge Engineering (ICCKE)**. Mashhad, Iran, Islamic Republic of, 2023. p. 636–641.