



UNIVERSIDADE FEDERAL DE SERGIPE
CENTRO DE CIÊNCIAS EXATAS E TECNOLOGIA
DEPARTAMENTO DE ENGENHARIA CIVIL

Luiz Antônio Muñoz Dantas

**Aplicação de redes neurais artificiais e do
algoritmo dos k -vizinhos mais próximos para
modelar o processo chuva-vazão na Bacia do Rio
Piauitinga (SE)**

São Cristóvão – SE

2022

Luiz Antônio Muñoz Dantas

**Aplicação de redes neurais artificiais e do algoritmo dos
 k -vizinhos mais próximos para modelar o processo
chuva-vazão na Bacia do Rio Piauitinga (SE)**

Trabalho de Conclusão de Curso apresentado ao
Departamento de Engenharia Civil da Univer-
sidade Federal de Sergipe (UFS) como requisito
para o título de Bacharel em Engenharia Civil

Universidade Federal de Sergipe (UFS)
Centro de Ciências Exatas e Tecnologia (CCET)
Departamento de Engenharia Civil (DEC)

Orientador: Prof. Dr. Alcigeimes Batista Celeste

São Cristóvão – SE

2022

Luiz Antônio Muñoz Dantas.

Aplicação de redes neurais artificiais e do algoritmo dos k -vizinhos mais próximos para modelar o processo chuva-vazão na Bacia do Rio Piauitinga (SE)/ Luiz Antônio Muñoz Dantas, São Cristóvão – SE, 2022. 48 p. : il.

Trabalho de Conclusão de Curso. Universidade Federal de Sergipe, São Cristóvão – SE.

I. Universidade Federal de Sergipe. II. Título.



UNIVERSIDADE FEDERAL DE SERGIPE
CENTRO DE CIÊNCIAS EXATAS E TECNOLOGIA
DEPARTAMENTO DE ENGENHARIA CIVIL



ATA DE DEFESA

Luiz Antônio Muñoz Dantas

Aplicação de redes neurais artificiais e do algoritmo dos k -vizinhos mais próximos para modelar o processo chuva-vazão na Bacia do Rio Piauitinga (SE)

Trabalho de Conclusão de Curso apresentado ao Departamento de Engenharia Civil da Universidade Federal de Sergipe (UFS) como requisito para o título de Bacharel em Engenharia Civil.

Aprovada em: 17 de novembro de 2022

Banca Examinadora	Nota
Orientador: Prof. Dr. Alcigeimes Batista Celeste (UFS)	<u>10,0</u>
Examinador: Prof. Dr. Ludmilson Abritta Mendes (UFS)	<u>10,0</u>
Examinador: Eng. Civil Breno Lucas Oliveira Menezes (Mestrando PROEC/UFS)	<u>10,0</u>
Média Final:	<u>10,0</u>



Documento assinado digitalmente
ALCIGEIMES BATISTA CELESTE
Data: 30/11/2022 14:35:19-0300
Verifique em <https://verificador.iti.br>

Prof. Dr. Alcigeimes Batista Celeste (UFS)
Orientador

*Dedico este trabalho aos meus pais, que sempre acreditaram
e me apoiaram em todos os momentos da minha vida.*

Agradecimentos

Agradeço aos meus pais, Luiz Carlos e Fátima, por todo amor, apoio e dedicação proporcionados. Vocês foram essenciais para a formação dos meus valores.

Ao meu irmão, Gustavo, pela fraternidade, suporte e os divertidos momentos vividos.

Aos meus avós, Antônio, Efigênia (*in memoriam*), Gilson e Maria Hortencia, pelo carinho e atenção que sempre tiveram comigo.

Aos meus familiares, em especial a minha tia e madrinha, Maria Angélica, pelo afeto e assistência.

Aos amigos que tenho desde a escola, Helder, Lucas e Raymundo, espero que nossa amizade se perpetue pelo resto de nossas vidas.

A todos os colegas que fiz durante a graduação, Gabriel, Jamyson, Jonas, Josuel, Matheus, Thadeu e Pedro, vocês tornaram a caminhada muito mais leve.

Por fim, ao orientador, Prof. Alcigeimes, pelos ensinamentos transmitidos, pela paciência e por me apresentar esse campo de estudo incrível do aprendizado de máquina.

"Empty your mind. Be formless, shapeless, like water. Put water into a cup, it becomes the cup. Put water into a teapot, it becomes the teapot. Water can flow or creep or drip or crash. Be water, my friend."

(Bruce Lee)

Resumo

Este trabalho apresenta a aplicação de redes neurais artificiais e do algoritmo dos k -vizinhos mais próximos para modelar o processo chuva-vazão na bacia hidrográfica do rio Piauitinga, em Sergipe. Os dados de entrada utilizados na modelagem foram os do posto fluviométrico de Estância e do pluviométrico de Salgado, os quais foram divididos entre calibração e validação, selecionados de maneira randômica. A execução das simulações ocorreu através do software de aprendizado de máquina Weka, e foram testadas quatro formulações diferentes para as situações diária e mensal. A eficiência das simulações realizadas foi verificada a partir do coeficiente de Nash-Sutcliffe, conforme o qual obtiveram-se resultados consideráveis para as situações apresentadas, principalmente nos modelos baseados em redes neurais artificiais. A formulação contendo como atributos chuva e vazão de até três períodos anteriores foi a que apresentou os melhores resultados, atingindo índices de eficiência classificados como bons para a abordagem diária e muito bons para a mensal. Modelagens alternativas foram realizadas para comparar o modelo usando apenas um posto de chuva com outra usando a chuva média na bacia, e notou-se que o modelo que utiliza apenas o posto de precipitação de Salgado apresentou resultados melhores ou iguais aos demais. Por fim, foram feitas novas simulações que utilizaram os períodos iniciais para calibração e os finais para validação, e percebeu-se que as simulações que classificaram os dados de forma randômica apresentaram resultados bastante superiores.

Palavras-chaves: Modelagem Chuva-Vazão. Redes Neurais Artificiais. Algoritmo dos k -vizinhos mais Próximos. Weka.

Abstract

This work presents the application of artificial neural networks and the k -nearest neighbors algorithm to model the rainfall-runoff process in the Piauitinga river basin, Sergipe. The input data used in the modeling were those from the Estância streamgauge station and from the Salgado raingauge station, which were grouped into calibration and validation sets, selected randomly. The simulation runs took place by means of the machine learning software Weka, and four different formulations were tested for the daily and monthly situations. The efficiency of the simulations performed was verified by the Nash-Sutcliffe coefficient, according to which considerable results were obtained for the situations presented, mainly for the models based on artificial neural networks. The formulation containing as features rainfall and streamflow of up to three previous periods was the one with best results, providing efficiency classified as good for the daily approach and very good for the monthly approach. Alternative modeling was carried out to compare the model using only one raingauge station with another using the average rainfall in the basin, and it was noted that the model that uses only the Salgado raingauge station achieved better or equal results than the others. Finally, new simulations were carried out using the initial periods for calibration and the final ones for validation, and it was noticed that the simulations that randomly classified the data gave much better results.

Keywords: Rainfall-runoff Modeling. Artificial Neural Networks. k -Nearest Neighbors. Weka.

Lista de ilustrações

Figura 1	– Rede neural artificial do tipo <i>feedforward</i>	18
Figura 2	– Detalhes de um neurônio artificial na camada oculta	18
Figura 3	– Gráfico da função sigmóide	19
Figura 4	– Menu do Weka	21
Figura 5	– Bacias hidrográficas de Sergipe com marcações indicando as localizações aproximadas dos postos utilizados	22
Figura 6	– Exemplo da configuração de um arquivo ARFF	25
Figura 7	– Captura de tela da aba <i>Preprocess</i>	25
Figura 8	– Captura de tela da localização da RNA no Weka	26
Figura 9	– Captura de tela da localização do KNN no Weka	27
Figura 10	– Exemplo de resultados simulados pelo WEKA	28
Figura 11	– Eficiências de Nash-Sutcliffe na Formulação 1 com aplicação de RNAs . . .	30
Figura 12	– Eficiências de Nash-Sutcliffe na Formulação 2 com aplicação de RNAs . . .	31
Figura 13	– Eficiências de Nash-Sutcliffe na Formulação 3 com aplicação de RNAs . . .	31
Figura 14	– Eficiências de Nash-Sutcliffe na Formulação 4 com aplicação de RNAs . . .	31
Figura 15	– Eficiências de Nash-Sutcliffe na Formulação 1 com aplicação do KNN . . .	32
Figura 16	– Eficiências de Nash-Sutcliffe na Formulação 2 com aplicação do KNN . . .	32
Figura 17	– Eficiências de Nash-Sutcliffe na Formulação 3 com aplicação do KNN . . .	32
Figura 18	– Eficiências de Nash-Sutcliffe na Formulação 4 com aplicação do KNN . . .	33
Figura 19	– Melhores índices de desempenho nas Formulações	34
Figura 20	– Eficiências de Nash-Sutcliffe para modelagem com diferentes dados de precipitação	35
Figura 21	– Influência da seleção dos dados de calibração e validação	36

Lista de tabelas

Tabela 1	–	Estações hidrometeorológicas utilizadas	23
Tabela 2	–	Formulações a serem investigadas nos DDM	24
Tabela 3	–	Configurações das camadas ocultas na RNA para as formulações 1 e 2 . . .	26
Tabela 4	–	Configurações das camadas ocultas na RNA para as formulações 3 e 4 . . .	27
Tabela 5	–	Configurações da quantidade de k -vizinhos utilizados no KNN	27

Lista de abreviaturas e siglas

ANA	Agência Nacional de Águas e Saneamento Básico
DDM	Modelo guiado por dados
IBL	Aprendizado baseado em instâncias
KNN	Algoritmo dos k -vizinhos mais próximos
RNA	Rede neural artificial
MLP	<i>Multilayer Perceptron</i>

Sumário

1	INTRODUÇÃO	14
1.1	Considerações Iniciais	14
1.2	Objetivos	15
1.2.1	Objetivo Geral	15
1.2.2	Objetivos específicos	15
2	REVISÃO BIBLIOGRÁFICA	16
2.1	Modelos Chuva-Vazão	16
2.2	Redes Neurais Artificiais	17
2.3	Algoritmo dos <i>k</i>-vizinhos mais Próximos	19
2.4	Eficiência de Nash-Sutcliffe	20
2.5	Software Weka para Modelagem Guiada por Dados	21
3	MATERIAL E MÉTODOS	22
3.1	Área de Estudo	22
3.2	Dados de Precipitação e Vazão	22
3.2.1	Obtenção dos Dados	22
3.2.2	Formulação dos DDM	23
3.2.3	Classificação dos Dados	23
3.3	Aplicação do Weka	24
3.3.1	Criação do arquivo ARFF	24
3.3.2	Seleção e Configuração do DDM	24
3.3.2.1	Configuração da RNA	24
3.3.2.2	Configuração do KNN	27
3.3.3	Simulação dos Modelos	27
3.4	Eficiência da Modelagem	28
4	RESULTADOS E DISCUSSÃO	30
4.1	Apresentação dos Resultados	30
4.1.1	Resultados da RNA	30
4.1.2	Resultados do KNN	30
4.2	Discussão	33
4.2.1	Melhores Configurações	33
4.2.2	Comparação entre os postos de precipitação	34
4.2.3	Influência da seleção dos dados de calibração e validação	36

Conclusão	37
REFERÊNCIAS	39
APÊNDICES	40
APÊNDICE A – EXEMPLO DE VISUALIZAÇÃO GERADA PELO WEKA PARA AS RNAS	41
APÊNDICE B – EXEMPLO DE PESOS E FUNÇÕES DE ATIVAÇÃO UTILIZADOS DURANTE O PROCESSO DE TREINAMENTO DE UMA RNA NO WEKA	42
APÊNDICE C – EFICIÊNCIAS DE NASH-SUTCLIFFE NAS FORMULAÇÕES 1 E 2 PARA A ABORDAGEM DIÁRIA COM RNA	43
APÊNDICE D – EFICIÊNCIAS DE NASH-SUTCLIFFE NAS FORMULAÇÕES 1 E 2 PARA A ABORDAGEM MENSAL COM RNA	44
APÊNDICE E – EFICIÊNCIAS DE NASH-SUTCLIFFE NAS FORMULAÇÕES 3 E 4 PARA A ABORDAGEM DIÁRIA COM RNA	45
APÊNDICE F – EFICIÊNCIAS DE NASH-SUTCLIFFE NAS FORMULAÇÕES 3 E 4 PARA A ABORDAGEM MENSAL COM RNA	46
APÊNDICE G – EFICIÊNCIAS DE NASH-SUTCLIFFE NAS FORMULAÇÕES 1 E 2 PARA O KNN	47
APÊNDICE H – EFICIÊNCIAS DE NASH-SUTCLIFFE NAS FORMULAÇÕES 3 E 4 PARA O KNN	48

1 Introdução

1.1 Considerações Iniciais

A água é um elemento fundamental para manutenção da vida, do meio ambiente e que está presente em diversas atividades humanas. O Brasil é um país rico em disponibilidade hídrica, contudo cerca de 70% da água doce disponível encontra-se na região amazônica, que contempla apenas 5% da população do país (SETTI et al., 2001).

A quantidade e a qualidade de água está essencialmente relacionada ao desenvolvimento social e econômico das regiões. Porém, os recursos hídricos estão distribuídos de forma desigual no planeta Terra, com isso existem regiões que apresentam escassez desse recurso e enfrentam problemas decorrentes das secas, e regiões que são abundantes em água e enfrentam cheias. Além disso, elementos como aumento das demandas, crescimento populacional e variações climáticas aumentam a complexidade do gerenciamento dos recursos hídricos (FARIAS; SANTOS; CELESTE, 2010).

Com o intuito de auxiliar na gestão adequada dos diferentes usos da água, como abastecimento, irrigação, geração de energia e indústria, além da previsão de eventos extremos de cheias e períodos de estiagem, é empregado comumente a modelagem chuva-vazão, baseada no complexo processo de transformação da precipitação em vazão (MACHADO, 2005). Muitos desses modelos já elaborados são do tipo conceitual, entretanto modelos empíricos do tipo ‘caixa-preta’, ou guiados por dados (DDM: *data driven models*) começaram a aparecer gradativamente na literatura nos últimos anos (CELESTE; MENEZES; FONTES, 2014).

Para a aplicação dos modelos conceituais é necessário a compreensão dos processos hidrológicos e exige-se profundo conhecimento das características físicas da bacia hidrográfica. Enquanto que para os DDMs, não é preciso ter conhecimento dessas características (SOUSA; SOUSA, 2006).

Existem diversos DDMs que podem ser empregados para solucionar o problema da modelagem chuva-vazão. As redes neurais artificiais (RNA) são um dos modelos mais populares (CELESTE; MENEZES; FONTES, 2014), e têm sido vastamente utilizadas para modelar sistemas não-lineares complexos com grande quantidade de dados, incluindo informações ruidosas (SOUSA; SOUSA, 2006). O aprendizado baseado em instância (IBL: *instance-based learning*) é uma outra técnica que também pode ser empregada para realização da modelagem chuva-vazão, e sua forma mais popular é o modelo do algoritmo dos k -vizinhos mais próximos (KNN) (AKBARI; OVERLOOP; AFSHAR, 2011).

Neste estudo, foi proposta a aplicação de DDMs para modelar o processo chuva-vazão na bacia hidrográfica do rio Piauitinga, em Sergipe, através de um software de aprendizado de máquina chamado Weka (<<https://www.cs.waikato.ac.nz/ml/weka/>>).

1.2 Objetivos

1.2.1 Objetivo Geral

O objetivo geral do presente trabalho é aplicar a modelagem chuva-vazão utilizando as metodologias de redes neurais artificiais e do algoritmo dos k -vizinhos mais próximos na bacia hidrográfica do rio Piauitinga, localizada no estado de Sergipe.

1.2.2 Objetivos específicos

- Verificar a eficiência das simulações realizadas a partir dos DDMs (RNA e KNN) para as situações diária e mensal na bacia hidrográfica do rio Piauitinga;
- Analisar os resultados obtidos nas formulações aplicadas aos modelos guiados por dados;
- Comparar a aplicação do modelo usando apenas um posto de chuva com outra usando a chuva média na bacia;
- Investigar a influência da escolha randômica da classificação dos dados de calibração e validação.

2 Revisão Bibliográfica

2.1 Modelos Chuva-Vazão

A modelagem chuva-vazão é uma técnica baseada na análise do comportamento da vazão em uma seção de rio de uma bacia hidrográfica em função da precipitação na bacia. Sua aplicação pode ser empregada para diferentes propósitos, como previsão de cheias e de aflúências em reservatórios (SCHEIDT; BRUNETTO, 2011).

Esse tipo de modelagem trata-se de um problema não linear, e existem diferentes maneiras aproximadas de solucioná-lo. Uma das formas de resolvê-lo é a partir de modelos conceituais, que levam em consideração as características físicas da bacia hidrográfica, e informações sobre os parâmetros hidrológicos. Outro modo, é a partir de modelos empíricos, guiados por dados (DDM: *data-driven models*) que baseiam-se na análise e correlação dos dados de entrada e saída por meio de funções, que não necessariamente estão relacionadas aos processos físicos envolvidos (CELESTE; MENEZES; FONTES, 2014; SOLOMATINE; MASKEY; SHRESTHA, 2008). Na literatura existem diversos exemplos de trabalhos relacionados à modelagem chuva-vazão, em que podem-se encontrar aplicações de modelos conceituais e empíricos.

Celeste e Chaves (2014) utilizaram o modelo conceitual *Tank Model* para simular o processo de chuva-vazão na bacia hidrográfica do rio Japaratuba, em Sergipe. Os autores analisaram diferentes algoritmos otimizadores e funções objetivo para calibração automática dos parâmetros do modelo. Os resultados obtidos foram comparados com uma pesquisa anterior, que utilizou uma combinação única de otimizador e função objetivo, e com o modelo conceitual MODHAC (VIEGAS FILHO et al., 1999), utilizado no estudo de elaboração do Plano Estadual de Recursos Hídricos do estado de Sergipe. Assim, concluíram que os parâmetros do modelo podem ser calibrados por diversos arranjos de otimizadores e funções objetivo, e foram alcançados resultados mais eficientes do que os obtidos no estudo anterior e no MODHAC.

Celeste, Menezes e Fontes (2014) fizeram a utilização do modelo Tank Model e da modelagem empírica baseada em redes neurais artificiais (RNA), para também modelar o processo de chuva-vazão na bacia hidrográfica do rio Japaratuba, em Sergipe. Neste estudo, os resultados obtidos foram confrontados entre si e também com o modelo conceitual MODHAC. Por fim, ao analisar os resultados dos onze modelos de RNAs com estrutura *feedforward*, sete tiveram desempenho superior ao Tank Model, e quatro tiveram desempenho inferior. Além disso, tanto o Tank Model quanto todos os onze modelos baseados em RNAs apresentaram melhores ajustes em relação ao modelo MODHAC.

Sousa e Sousa (2006) realizaram a aplicação de RNA do tipo *Multilayer Perceptron* (MLP) para modelar o processo de chuva-vazão na bacia hidrográfica de Caraúbas, na Paraíba. Os autores tinham como finalidade prever a vazão média mensal na seção transversal do rio Paraíba,

a montate de Caraúbas, a partir da modelagem com RNA. Para isso, treinaram três diferentes arquiteturas de RNA com o intuito de encontrar a que apresentasse a melhor previsão. Por fim, selecionaram a arquitetura que considerou seis camadas escondidas, representadas pelo módulo da chuva dos postos de Caraúbas, Monteiro, Congo, São Sebastião do Umbuzeiro e Camalaú, e uma camada de saída representada pelo módulo da vazão, e concluíram que a RNA apresentou bons resultados de previsão de vazão média mensal, mas não foi capaz de prever adequadamente os picos de cheia.

Scheidt e Brunetto (2011) utilizaram RNAs acopladas a um algoritmo genético para modelagem do processo chuva-vazão de uma bacia hidrográfica do rio Piquiri, no Paraná. Neste trabalho, compararam-se os resultados de uma modelagem aplicando apenas uma RNA do tipo MLP e outra utilizando uma RNA MLP com algoritmo genético. Os autores concluíram que, ao confrontar os resultados, a utilização de algoritmos genéticos proporcionou um melhor desempenho da modelagem. Contudo, advertiram que esse modelo pode aumentar consideravelmente o tempo de processamento, a depender da quantidade de dados de treinamento.

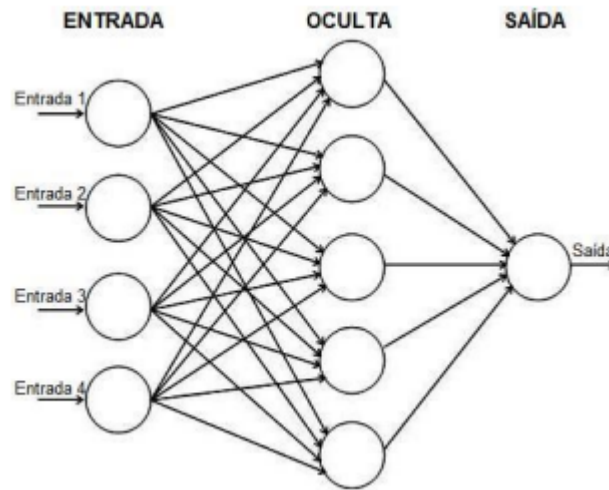
2.2 Redes Neurais Artificiais

A RNA é um modelo empírico que baseia-se no funcionamento do cérebro humano, através de neurônios artificiais que recebem e enviam informações semelhantes às células do tecido nervoso animal (MACHADO, 2005).

Existem diferentes arquiteturas de RNA que podem ser adotadas e neste trabalho é utilizado uma do tipo *feedforward* (Figura 1), constituída por camadas de entrada, intermediária(s) (ou oculta(s)) e de saída. Neste tipo de RNA, a quantidade de neurônios na camada de entrada altera-se de acordo com a quantidade de variáveis de entrada inseridas. Na camada oculta pode-se obter essa quantidade a partir da comparação de testes com diferentes configurações, e seleção da que apresentar o melhor resultado. A camada de saída é constituída por apenas um neurônio, que retorna o dado de saída calculado (CELESTE; MENEZES; FONTES, 2014).

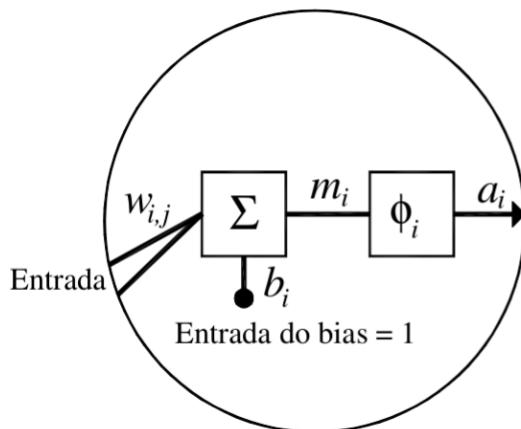
Nesse tipo de configuração não há conexão direta entre os neurônios da camada de entrada e os da camada de saída. As conexões ocorrem de maneira que os neurônios da camada de entrada se conectam aos da camada oculta, e os neurônios da camada intermediária conectam-se com os da camada de saída, o que faz com que os dados percorram uma única direção através das camadas. Entre as conexões existem pesos que são calibrados durante o processo de treinamento, e servem para reduzir os erros na saída da RNA (MACHADO, 2005).

A partir do processo de recebimento de dados, processamento e geração de uma saída, os neurônios permitem que as RNAs possam conceber relações não lineares entre os dados de entrada e de saída. Na Figura 2 pode-se observar a configuração de um neurônio artificial inserido em uma camada oculta, em que cada elemento de entrada conecta-se ao neurônio. Com isso, ocorre o somatório da multiplicação dos dados de entrada pelo conjunto de pesos (w_{ij}) e do *bias* (b_i), gerando assim o campo local induzido (m_i). Por último, para que a informação seja

Figura 1 – Rede neural artificial do tipo *feedforward*

Fonte: Celeste, Menezes e Fontes (2014).

Figura 2 – Detalhes de um neurônio artificial na camada oculta



onde,

$w_{i,j}$ = peso (i indica o neurônio de destino do peso e j indica qual é o neurônio de partida deste peso);

b_i = bias do i -ésimo neurônio;

m_i = campo local induzido do i -ésimo neurônio;

ϕ_i = função de ativação do i -ésimo neurônio;

a_i = saída do i -ésimo neurônio.

Fonte: Farias, Santos e Celeste (2010).

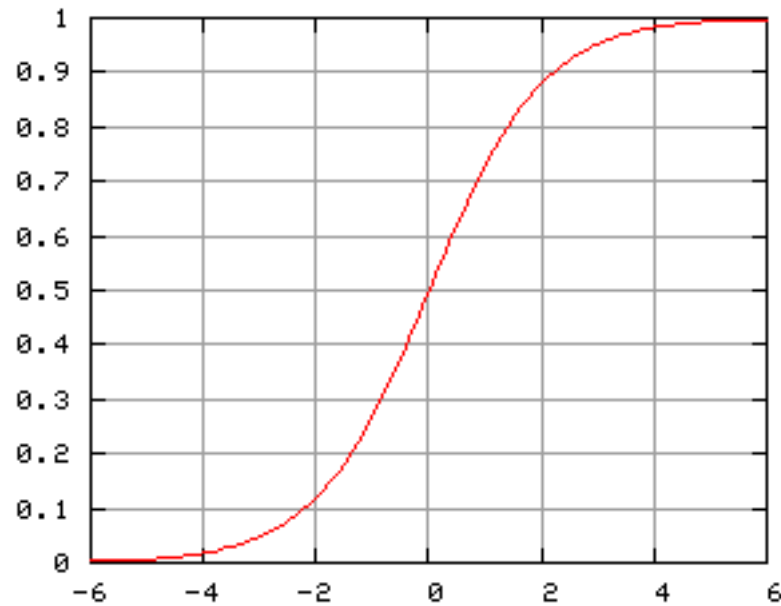
transferida para os neurônios adjacentes, o campo local induzido é submetido a uma função de ativação ϕ_i , obtendo-se assim o valor de entrada do neurônio subsequente (CELESTE; MENEZES; FONTES, 2014).

A função de ativação (ou de transferência) possibilita que a RNA identifique relações entre as variáveis de entrada e saída (MACHADO, 2005). Existem diversos tipos de funções que podem ser adotadas como função de transferência, e neste estudo foram utilizadas funções do tipo sigmóides, na saída dos neurônios das camadas ocultas, e do tipo linear na saída do único neurônio da camada de saída.

A função sigmóide pode ser definida a partir da equação 2.1, e seu gráfico é representado pela Figura 3.

$$f(x) = \frac{1}{1 + e^{-x}} \quad (2.1)$$

Figura 3 – Gráfico da função sigmóide



Fonte: CC BY-SA 3.0, <https://commons.wikimedia.org/w/index.php?curid=644559>.

2.3 Algoritmo dos k -vizinhos mais Próximos

Um dos problemas envolvendo DDMs é a questão da transparência, pois muitas vezes sofrem por ficarem encapsuladas em códigos de software, e dificultam assim a aceitação do usuário final. Dessa forma, uma das opções para essa problemática é a utilização de uma outra técnica de *machine learning*, conhecida como aprendizado baseado em instâncias (IBL: *instance-based learning*) que, com base na combinação dos dados existentes (instâncias, e.g., registros históricos, resultados de um experimento, etc.) obtém a predição (SOLOMATINE; MASKEY; SHRESTHA, 2008).

A RNA utiliza uma técnica conhecida como *eager learning* que encontra uma relação explícita entre os dados, e necessita de uma fase de treinamento. Em contraste, o IBL baseia-se em um método chamado *lazy learning*, que consiste em armazenar os dados de treinamento e, quando uma nova entrada é fornecida, busca-se na memória um conjunto de instâncias relacionadas e suas saídas são utilizadas para obter o resultado da nova saída (SOLOMATINE; MASKEY; SHRESTHA, 2008).

De acordo com Akbari, Overloop e Afshar (2011), um dos modelos mais populares que utilizam o IBL é o *k-nearest neighbor* (KNN), uma versão do algoritmo dos vizinhos mais próximos (*nearest neighbor*), que classifica um novo exemplo de consulta (*query vector* X_q) a partir da combinação dos valores dos k vizinhos selecionados. Assim, uma norma Euclidiana ponderada é geralmente utilizada para medir a proximidade (similaridade) entre os vetores X_i e X_q , respectivamente, vetor de atributos e vetor de consulta (não incluído no conjunto de treinamento):

$$d(X_q, X_i) = \sqrt{\sum_{j=1}^n w_j^A (x_{ij} - x_{qj})^2} \quad (2.2)$$

onde w_j^A é o peso do j -ésimo atributo. Com os valores de saída y_i dos k pontos mais próximos (vizinhos), o valor previsto para y_q é calculado pela média ponderada:

$$y_q = \frac{\sum_{i=1}^k w_i^V y_i}{\sum_{i=1}^k w_i^V} \quad (2.3)$$

onde w_i^V é o peso do i -ésimo vizinho. Em sua versão mais simples, $w_i^V = 1/k$ e a estimativa é o valor médio de k vizinhos mais próximos. Todavia, em sua versão completa, cada vizinho recebe um peso de acordo com a distância $d(X_q, X_i)$, em que um vizinho mais distante recebe um peso menor, exercendo menor influência do que os vizinhos que estão mais próximos. Sendo assim, algumas funções *kernel* que decrescem monotonicamente à medida que a distância aumenta, têm sido utilizadas (AKBARI; OVERLOOP; AFSHAR, 2011):

$$\text{Linear: } w_i^V = 1 - d(X_q, X_i)$$

$$\text{Inversa: } w_i^V = [d(X_q, X_i)]^{-1}$$

$$\text{Inversa quadrática: } w_i^V = \{[d(X_q, X_i)]^2\}^{-1}$$

$$\text{Exponencial: } w_i^V = \exp[-d(X_q, X_i)]$$

$$\text{Gaussiana: } w_i^V = \exp\{-[d(X_q, X_i)]^2\}$$

Contudo, de acordo com Atkeson (1997, apud Akbari, Overloop e Afshar, 2011) não existem evidências que uma função kernel específica seja superior às outras.

Ademais, segundo Kang e Cho (2008, apud Akbari, Overloop e Afshar, 2011) o número de vizinhos é muitas vezes escolhido empiricamente, por validação cruzada ou pelo conhecimento prático. Portanto, não existe um método completamente estabelecido para seleção da quantidade ideal de vizinhos utilizados no KNN.

2.4 Eficiência de Nash-Sutcliffe

Neste trabalho, para realizar a comparação entre as vazões observadas e as previstas pelos modelos, utilizou-se como medida de desempenho a eficiência de Nash-Sutcliffe (E), conforme a equação abaixo:

$$E = 1 - \frac{\sum_{i=1}^m (y_i - Y_i)^2}{\sum_{i=1}^T (y_i - \bar{y})^2} \quad (2.4)$$

onde, y_i é o valor observado, Y_i é o valor previsto, $\bar{y}$ é a média dos valores observados e m é o número de amostras (instâncias).

A eficiência de Nash-Sutcliffe pode variar de $-\infty$ a 1, em que quanto mais próximo ao valor unitário, melhor é o modelo.

Uma eficiência $E = 1$ indica o ajuste perfeito do modelo aos dados observados. Uma eficiência $E = 0$ quer dizer que as previsões do modelo são tão precisas quanto a média dos

Figura 4 – Menu do Weka



Fonte: O autor.

dados observados. Por fim, uma eficiência inferior a zero significa que a média observada é um indicador melhor do que o modelo (CELESTE; CHAVES, 2014).

2.5 Software Weka para Modelagem Guiada por Dados

O Weka (*Waikato Environment for Knowledge Analysis*) é um *software* desenvolvido na Universidade de Waikato (Nova Zelândia), multiplataforma e escrito na linguagem Java. O programa disponibiliza uma coleção de algoritmos de aprendizagem de máquina e ferramentas de pré-processamento de dados. Foi projetado para que os usuários possam rapidamente experimentar os métodos existentes em novos conjuntos de dados (FRANK; HALL; WITTEN, 2016).

Assim, através do Weka é possível facilmente aplicar técnicas de *machine learning* sem a necessidade de escrever códigos de programação. O presente trabalho utilizou esse software para a aplicação de RNA e KNN na modelagem chuva-vazão. A interface inicial do programa pode ser observada na Figura 4.

3 Material e Métodos

3.1 Área de Estudo

Neste estudo será empregado a modelagem chuva-vazão na bacia hidrográfica do Rio Piauitinga, em Sergipe, que integra a bacia hidrográfica do Rio Piauí (Figura 5).

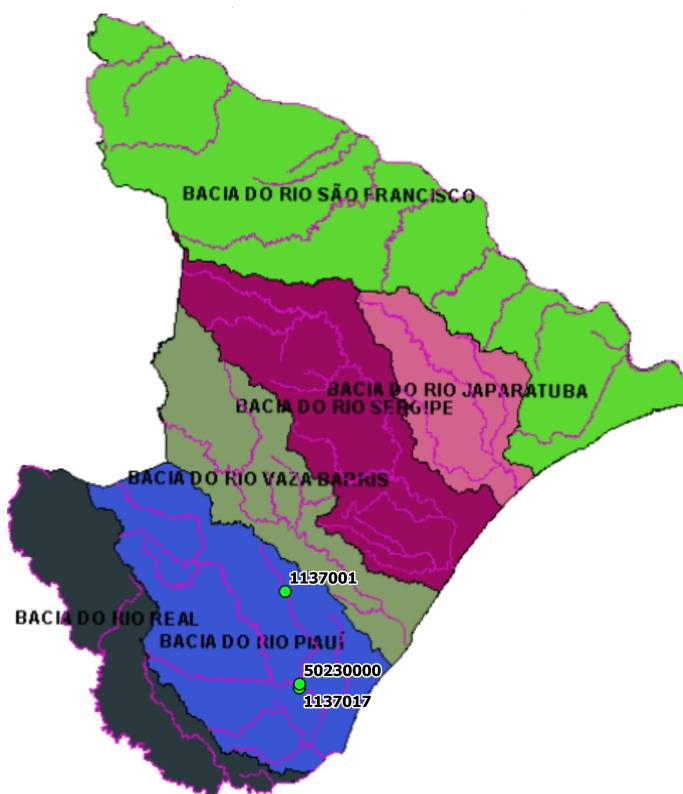
De acordo com Lima (2022), a bacia do Rio Piauitinga possui uma área de 417,42 km², e seu principal rio possui uma extensão de 59,11 km. Além disso, o rio Piauitinga é afluente da margem esquerda do rio Piauí, e sua confluência ocorre na cidade de Estância.

3.2 Dados de Precipitação e Vazão

3.2.1 Obtenção dos Dados

O levantamento de dados de precipitação e vazão é uma etapa fundamental para modelagem chuva-vazão, pois constituem os atributos de entrada e os dados de saída observados.

Figura 5 – Bacias hidrográficas de Sergipe com marcações indicando as localizações aproximadas dos postos utilizados



Fonte: Adaptado de SEMARH-SE (2012).

Tabela 1 – Estações hidrometeorológicas utilizadas

Nome do Posto	Tipo	Código
Estância	Fluviométrico	50230000
Estância	Pluviométrico	1137017
Salgado	Pluviométrico	1137001

Fonte: O autor.

Neste estudo serão utilizados os dados disponíveis no portal HidroWeb, uma ferramenta da Agência Nacional de Águas e Saneamento Básico (ANA), que disponibiliza os dados coletados pelas estações hidrometeorológicas.

Dessa maneira, através do portal foi possível ter acesso às séries históricas de vazão do posto fluviométrico de Estância, e de chuva do pluviométrico de Salgado. Esses postos foram selecionados baseados em suas localizações.

Neste trabalho, para a realização da modelagem foi decidido utilizar abordagens diária e mensal. Desse modo, para os dados diários, foi utilizado o período de 1 de janeiro de 1994 até 10 de junho de 2004 (10 anos e 5 meses), pois não foram registradas ocorrências de falhas. Já para os dados mensais, foi utilizado o período de janeiro de 1966 até junho de 2006 (40 anos e 5 meses). Contudo, no posto de Salgado foram verificadas falhas no intervalo de outubro de 1992 até março de 1993. Sendo assim, foi feita a substituição das informações faltosas pelos respectivos dados do posto pluviométrico de Estância. Os postos utilizados podem ser observados na Tabela 1.

3.2.2 Formulação dos DDM

Para a realização da modelagem chuva-vazão, os DDMs foram submetidos a quatro tipos diferentes de formulações, da maneira:

$$h(x) = h(x_1, x_2, \dots, x_n) \quad (3.1)$$

onde, $h(x)$ é a hipótese para vazão $Q(t)$ do dia/mês t em função dos atributos $x_1, x_2, \dots, x_n$. Esses atributos são constituídos por precipitações do dia/mês atual ($P(t)$) ou de períodos passados ($P(t-1), P(t-2), \dots, (P(t-n))$) e/ou vazões de períodos passados ($Q(t-1), Q(t-2), \dots, Q(t-n)$).

Os dados de precipitação e de vazão que são importados através do portal HidroWeb, são dispostos de maneira matricial. Desse jeito, por meio de um editor virtual de planilhas, foi feita a transformação dos dados para a forma vetorial, posicionando-os em seus respectivos atributos de entrada e na variável de saída observada (vazão observada). As formulações dos DDMs foram escolhidas de maneira empírica, e podem ser observadas na Tabela 2.

3.2.3 Classificação dos Dados

Para o desenvolvimento da modelagem realizada pelos DDMs, os conjuntos de dados foram divididos entre calibração e validação.

Tabela 2 – Formulações a serem investigadas nos DDM

1	$Q(t) = h(Q(t-1), Q(t-2), Q(t-3), Q(t-4), Q(t-5))$
2	$Q(t) = h(P(t), P(t-1), P(t-2), P(t-3), P(t-4), P(t-5))$
3	$Q(t) = h(P(t), P(t-1), P(t-2), P(t-3), Q(t-1), Q(t-2), Q(t-3))$
4	$Q(t) = (P(t-1), P(t-2), P(t-3), P(t-4), Q(t-1), Q(t-2), Q(t-3), Q(t-4))$

Fonte: O autor.

Os dados de calibração são utilizados para treinar o modelo, enquanto que os dados de validação são utilizados para validar o modelo. Neste trabalho, foram utilizados 70% dos dados para calibração e 30% para validação, selecionados de maneira randômica, i.e., tantos dados de períodos passados quanto de presentes podem fazer parte dos dois conjuntos.

3.3 Aplicação do Weka

3.3.1 Criação do arquivo ARFF

Para a utilização do software Weka, é necessário realizar a exportação dos dados para o programa. Dessa maneira, é preciso converter os dados da planilha para o formato ARFF, que é o padrão utilizado pelo Weka.

Esse formato consiste em uma lista de instâncias, e os valores de atributo para cada instância são separados por vírgula. Além disso, o arquivo ARFF deve conter indicações (marcações) para que o programa realize a interpretação dos dados de maneira apropriada, como pode-se observar na Figura 6 (FRANK; HALL; WITTEN, 2016). Para configurar o arquivo utiliza-se um editor de texto, e após isso basta salvá-lo no formato ARFF.

Desse modo, para realização da modelagem foram criados dois tipos de arquivos ARFF, um denominado *trainData*, constituído pelo conjunto de dados de calibração, e outro chamado de *allData*, que abrange todo o conjunto de dados (calibração e validação).

Em seguida, para carregar o arquivo no Weka, deve-se selecionar a opção *Explorer* no menu inicial do programa (Figura 4). Posteriormente, basta abrir o arquivo *trainData* clicando em *Open File* na aba de *Preprocess*. Nessa aba, o programa possibilita visualizar informações relacionadas aos atributos (Figura 7), como algumas estatísticas básicas e histogramas.

3.3.2 Seleção e Configuração do DDM

Na sequência, na aba *Classify*, na opção *Choose*, é possível escolher qual DDM será utilizado. Como citado anteriormente, para esse estudo foram aplicados RNA e KNN.

3.3.2.1 Configuração da RNA

No WEKA, é possível trabalhar com RNA ao clicar na pasta *functions* e selecionar a opção *MultilayerPerceptron*, como observa-se na Figura 8.

Figura 6 – Exemplo da configuração de um arquivo ARFF

```

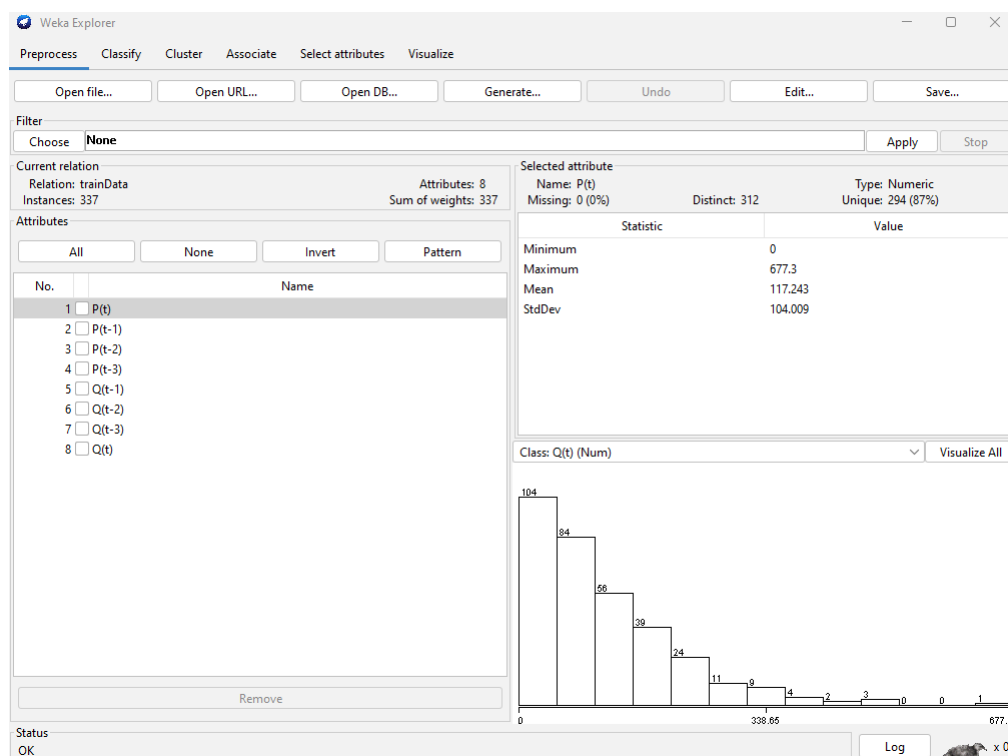
@RELATION trainData

@ATTRIBUTE P(t) NUMERIC
@ATTRIBUTE P(t-1) NUMERIC
@ATTRIBUTE P(t-2) NUMERIC
@ATTRIBUTE P(t-3) NUMERIC
@ATTRIBUTE Q(t-1) NUMERIC
@ATTRIBUTE Q(t-2) NUMERIC
@ATTRIBUTE Q(t-3) NUMERIC
@ATTRIBUTE Q(t) NUMERIC

@DATA
196.5,277.2,63.5,149.5,11,4.52,3.35,17
174.1,196.5,277.2,63.5,17,11,4.52,9.21
96.4,113.2,185.9,174.1,8.01,8.2,9.21,5.55
15.8,96.4,113.2,185.9,5.55,8.01,8.2,5.1
86.1,63,15.8,96.4,3.95,5.1,5.55,3.95
1.5,86.1,63,15.8,3.95,3.95,5.1,3.79
87,52.1,1.5,86.1,3.9,3.79,3.95,4.53
127.2,87,52.1,1.5,4.53,3.9,3.79,6.95
151.5,338.9,330.7,127.2,15,12.3,6.95,4.65
96.6,151.5,338.9,330.7,4.65,15,12.3,5.31
58.4,96.6,151.5,338.9,5.31,4.65,15,4.65
27.5,58.4,96.6,151.5,4.65,5.31,4.65,3.44
20.5,27.5,58.4,96.6,3.44,4.65,5.31,3.46
118.3,20.5,27.5,58.4,3.46,3.44,4.65,4.53
54.8,118.3,20.5,27.5,4.53,3.46,3.44,4.05
109.9,58.8,54.8,118.3,3.83,4.05,4.53,2.84
493.3,127.6,109.9,58.8,2.58,2.84,3.83,20.7
173.1,143.1,493.3,127.6,11.1,20.7,2.58,18.5
48.8,173.1,143.1,493.3,18.5,11.1,20.7,2.62
26.1,48.8,173.1,143.1,2.62,18.5,11.1,1.87
16.1,26.1,48.8,173.1,1.87,2.62,18.5,2.39
112,16.1,26.1,48.8,2.39,1.87,2.62,2.3

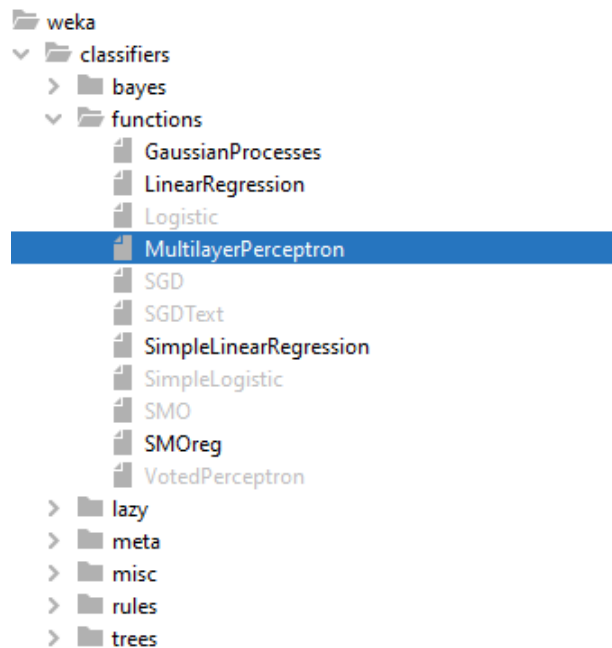
```

Fonte: O autor

Figura 7 – Captura de tela da aba *Preprocess*

Fonte: O autor.

Figura 8 – Captura de tela da localização da RNA no Weka



Fonte: O autor.

Tabela 3 – Configurações das camadas ocultas na RNA para as formulações 1 e 2

RNA-3	RNA-4	RNA-6
RNA-3,1	RNA-4,2	RNA-6,3
RNA-3,3	RNA-4,4	RNA-6,6
RNA-3,1,1	RNA-4,2,1	RNA-6,3,3
RNA-3,3,1	RNA-4,4,2	RNA-6,6,3
RNA-3,3,3	RNA-4,4,4	RNA-6,6,6

Fonte: O autor.

Além disso, baseado na composição da formulação utilizada, o Weka é capaz de gerar automaticamente uma configuração para a(s) camada(s) oculta(s) da RNA. Contudo, o programa também permite que o usuário configure sua própria topologia de RNA, pois possibilita a edição do número de neurônios e a quantidade de camadas ocultas. O software também disponibiliza a visualização das composições das RNAs utilizadas (Apêndice A).

O Weka sugeriu dois tipos de configurações diferentes para as camadas intermediárias das formulações investigadas (Tabela 2). Para as formulações 1 e 2, o programa utilizou três neurônios artificiais arranjados em uma única camada oculta. Já nas formulações 3 e 4, usou quatro neurônios, também dispostos em uma única camada.

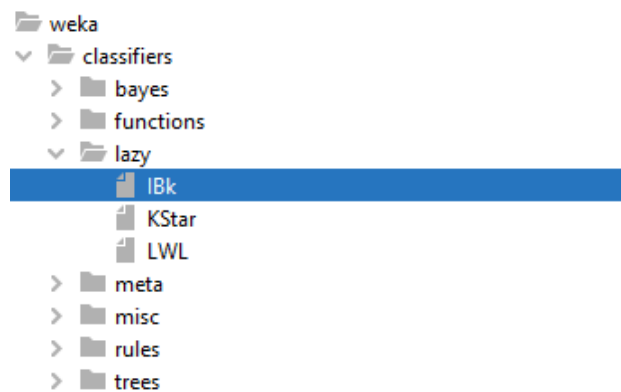
Dessa forma, neste trabalho foram testadas diversas configurações diferentes para RNA. Essas configurações foram escolhidas de maneira heurística, e definiu-se que para as formulações 1 e 2 seriam utilizadas as configurações da Tabela 3, e para as formulações 3 e 4 as da Tabela 4. Nesses esquemas, os valores numéricos significam a quantidade de neurônios que pertencem a uma camada oculta, e a adição de novas camadas é representada pela utilização de vírgulas.

Tabela 4 – Configurações das camadas ocultas na RNA para as formulações 3 e 4

RNA-4	RNA-6	RNA-8
RNA-4,2	RNA-6,3	RNA-8,4
RNA-4,4	RNA-6,6	RNA-8,8
RNA-4,2,1	RNA-6,3,3	RNA-8,4,4
RNA-4,4,2	RNA-6,6,3	RNA-8,8,4
RNA-4,4,4	RNA-6,6,6	RNA-8,8,8

Fonte: O autor.

Figura 9 – Captura de tela da localização do KNN no Weka



Fonte: O autor.

Tabela 5 – Configurações da quantidade de k -vizinhos utilizados no KNN

1-NN	7-NN	15-NN
3-NN	10-NN	17-NN
5-NN	13-NN	20-NN

Fonte: O autor.

3.3.2.2 Configuração do KNN

Para o KNN, é possível localizá-lo ao clicar na pasta *lazy* e escolher a opção *IBk*, como verifica-se na Figura 9.

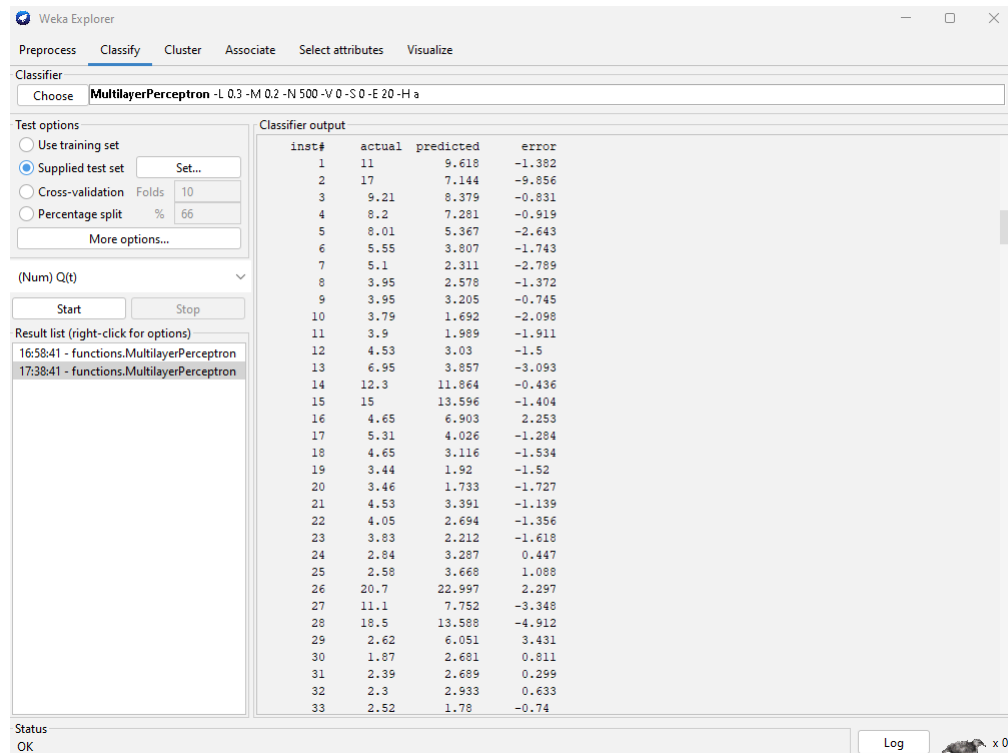
Desse modo, foram testadas simulações para o KNN como na Tabela 5, com o número de k -vizinhos escolhido de forma empírica. Neste trabalho foi aplicada a forma mais simplificada do KNN, sem fazer uso das funções *kernel*.

3.3.3 Simulação dos Modelos

Após selecionar o DDM que será aplicado, deve-se utilizar o arquivo que contém os dados de calibração (*trainData*) como conjunto de treinamento, a partir da opção *Use training set*. Com isso, o programa irá treinar o modelo. Para as simulações baseadas em RNAs, o software ainda fornece os pesos utilizados no processo de treinamento e as funções de ativação empregadas em cada neurônio (Apêndice B).

Em seguida, é inserido o conjunto de dados *allData*, contendo dados de calibração e

Figura 10 – Exemplo de resultados simulados pelo WEKA



Fonte: O autor.

validação, através da opção *Supplied test set*. Dessa forma, o Weka é submetido a um novo conjunto de dados (dados de validação), que não foram utilizados durante o processo de treinamento. Assim, pode-se testar a modelagem realizada pelo software.

Por último, a partir do arquivo *allData*, o programa fornece todos os resultados obtidos ao fim da simulação, tanto para os dados de calibração como os de validação. Na Figura 10, pode-se observar um exemplo disso, no qual o programa disponibiliza os resultados das vazões simuladas, além de informar a vazão observada e o erro entre elas.

3.4 Eficiência da Modelagem

Para avaliar o desempenho da simulação gerada pelo Weka, foi utilizada a eficiência de Nash-Sutcliffe (Equação 2.4). Desse modo, a equação foi empregada aos dados de calibração, dados de validação e ao conjunto completo de dados.

É interessante destacar os valores de Nash-Sutcliffe obtidos a partir dos dados de validação, pois estes não são utilizados durante o processo de treinamento do modelo. Sendo assim, contribuem para testar e ratificar a eficiência da modelagem desenvolvida.

Neste trabalho, os valores dos coeficientes de Nash-Sutcliffe encontrados pela equação 2.4 serão classificados de acordo com Baltokoski et al. (2010, apud Lima, 2022), da seguinte forma:

- Muito Bom: $E > 0,75$;

- Bom: $0,65 < E \leq 0,75$;
- Satisfatório: $0,50 < E \leq 0,65$;
- Insatisfatório: $E \leq 0,50$.

4 Resultados e Discussão

4.1 Apresentação dos Resultados

Para a apresentação dos valores de índice de desempenho, foram utilizados gráficos. Neles podem-se observar as configurações aplicadas aos modelos e os resultados obtidos para o coeficiente de Nash-Sutcliffe (considerando apenas os valores não negativos) para os dados de validação. Além disso, nos apêndices é possível verificar todos os coeficientes de eficiência que foram encontrados nas simulações para os dados de validação, calibração e completo.

4.1.1 Resultados da RNA

Utilizou-se as configurações das Tabelas 3 e 4 para obtenção dos resultados da RNA, conforme tratado no item 3.3.2.1.

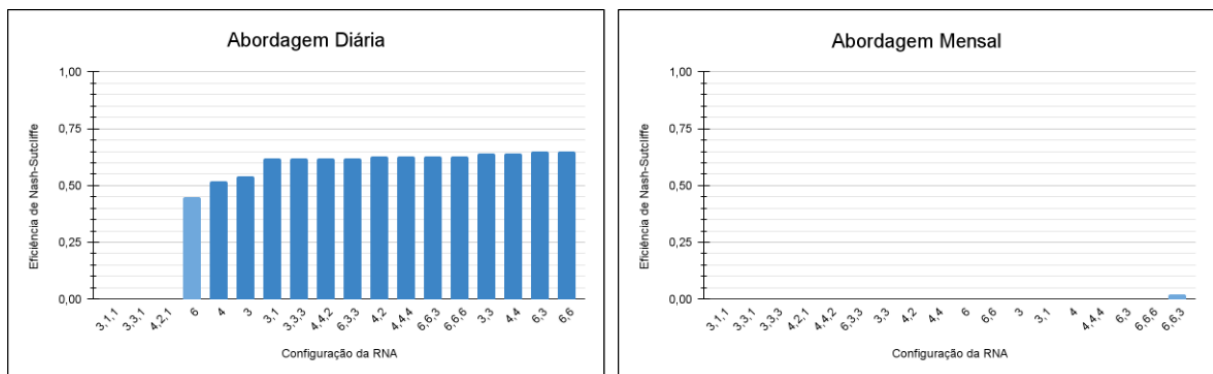
Os resultados de eficiência de Nash-Sutcliffe para os dados de validação nas situações diária e mensal, podem ser observados nas Figuras 11, 12, 13 e 14 para as Formulações 1, 2, 3 e 4, respectivamente. Além disso, nos Apêndices C e D podem-se verificar os coeficientes de desempenho para os dados de validação, calibração e completo encontrados nas Formulações 1 e 2. Já nos Apêndices E e F, os coeficientes obtidos nas Formulações 3 e 4.

4.1.2 Resultados do KNN

Para o KNN, foi utilizada a configuração da Tabela 5, como discutido no item 3.3.2.2.

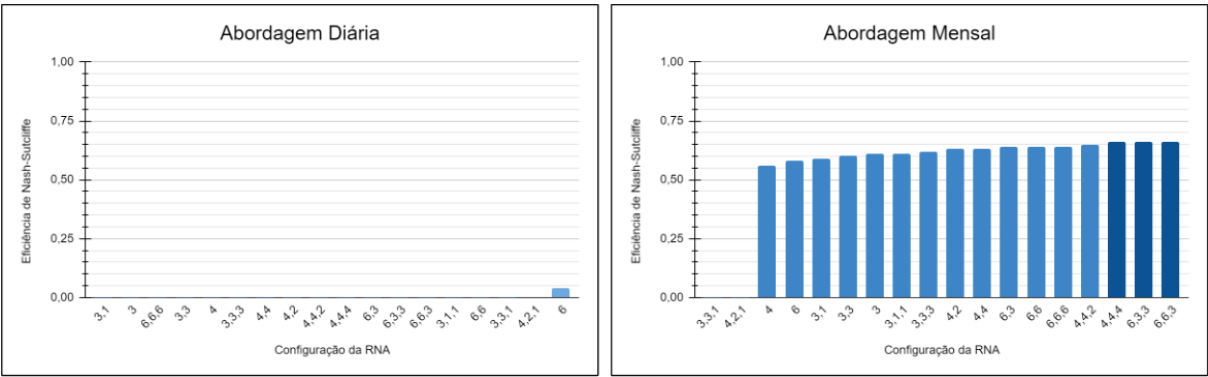
Os resultados de eficiência de Nash-Sutcliffe obtidos para os dados de validação nas Formulações 1, 2, 3 e 4 para as abordagens diária e mensal, podem ser encontrados nas Figuras 15, 16, 17 e 18, respectivamente. No Apêndice G, é possível observar os coeficientes de desempenho

Figura 11 – Eficiências de Nash-Sutcliffe na Formulação 1 com aplicação de RNAs



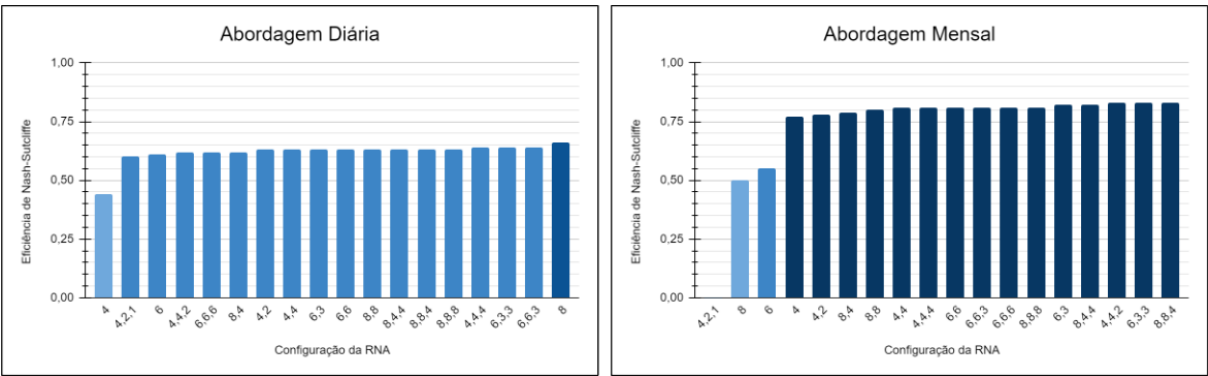
Fonte: O autor.

Figura 12 – Eficiências de Nash-Sutcliffe na Formulação 2 com aplicação de RNAs



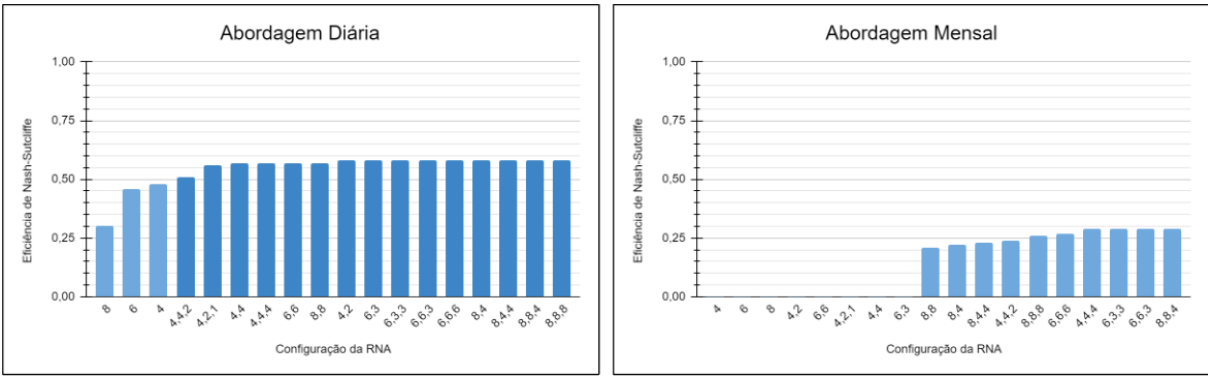
Fonte: O autor.

Figura 13 – Eficiências de Nash-Sutcliffe na Formulação 3 com aplicação de RNAs



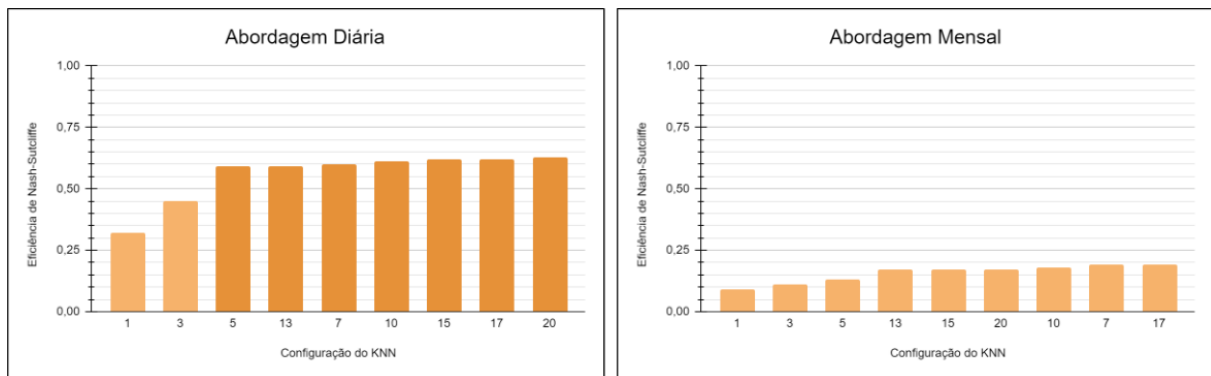
Fonte: O autor.

Figura 14 – Eficiências de Nash-Sutcliffe na Formulação 4 com aplicação de RNAs



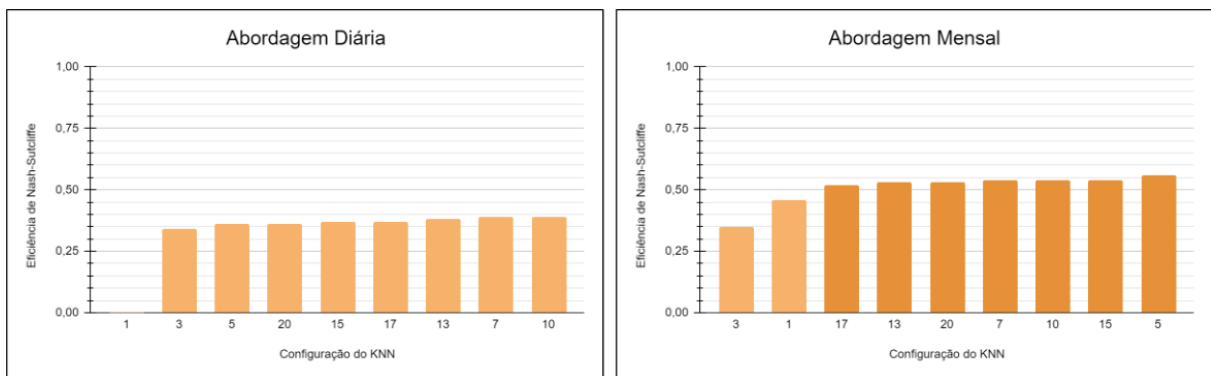
Fonte: O autor.

Figura 15 – Eficiências de Nash-Sutcliffe na Formulação 1 com aplicação do KNN



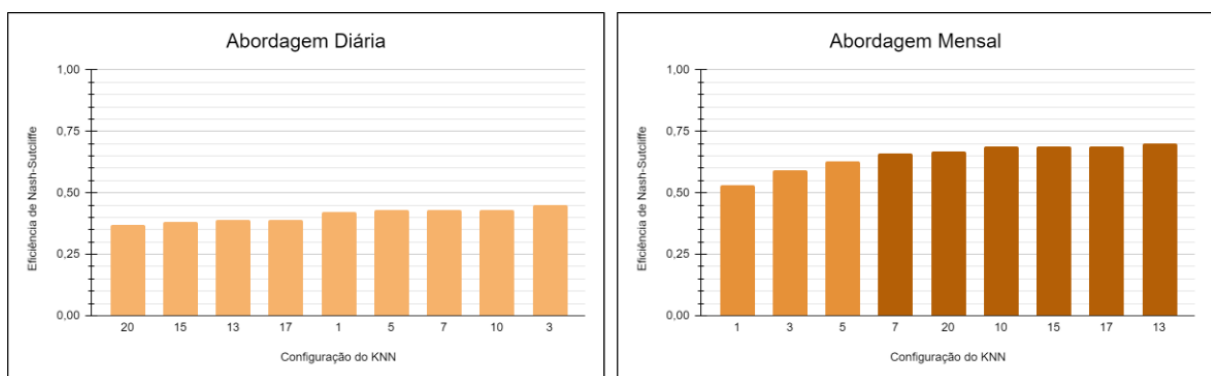
Fonte: O autor.

Figura 16 – Eficiências de Nash-Sutcliffe na Formulação 2 com aplicação do KNN



Fonte: O autor.

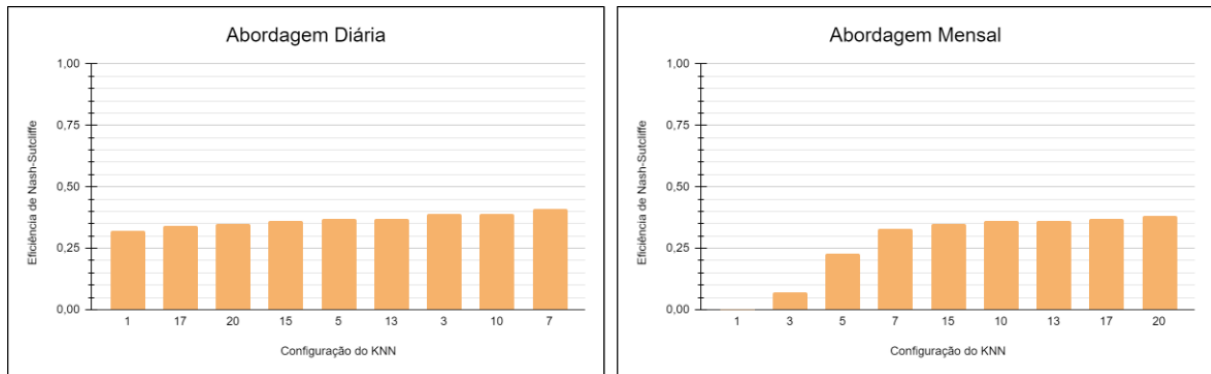
Figura 17 – Eficiências de Nash-Sutcliffe na Formulação 3 com aplicação do KNN



Fonte: O autor.

para os dados de validação, calibração e completo nas Formulações 1 e 2. Já no Apêndice H, os coeficientes referentes as Formulações 3 e 4.

Figura 18 – Eficiências de Nash-Sutcliffe na Formulação 4 com aplicação do KNN



Fonte: O autor.

4.2 Discussão

4.2.1 Melhores Configurações

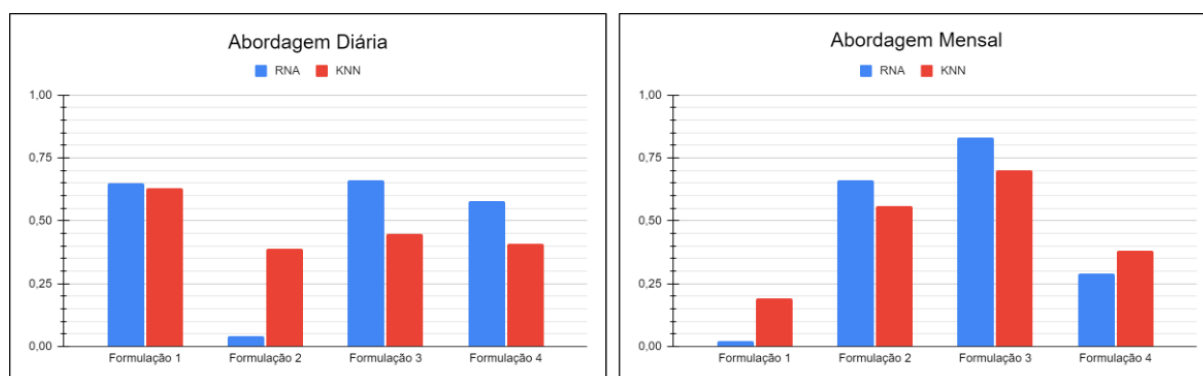
As configurações que apresentaram o melhor valor de coeficiente de Nash-Sutcliffe para os dados de validação, com a menor quantidade de neurônios e o menor número de k -vizinhos, são apresentadas abaixo:

- Abordagem diária:
 - Formulação 1: RNA-6,3 ($E = 0,65$); 20-NN ($E = 0,63$);
 - Formulação 2: RNA-6 ($E = 0,04$); 7-NN ($E = 0,39$);
 - Formulação 3: RNA-8 ($E = 0,66$); 3-NN ($E = 0,45$);
 - Formulação 4: RNA-4,2 ($E = 0,58$); 7-NN ($E = 0,41$).
- Abordagem mensal:
 - Formulação 1: RNA-6,6,3 ($E = 0,02$); 7-NN ($E = 0,19$);
 - Formulação 2: RNA-4,4,4 e RNA-6,3,3 ($E = 0,66$); 5-NN ($E = 0,56$);
 - Formulação 3: RNA-4,4,2 ($E = 0,83$); 13-NN ($E = 0,70$);
 - Formulação 4: RNA-4,4,4 e RNA-6,3,3 ($E = 0,29$); 20-NN ($E = 0,38$).

Na Figura 19 é possível observar os melhores índices de desempenho em cada Formulação. De modo geral é possível perceber que a formulação 3 foi a que mais se destacou. A formulação apresentou os melhores valores de coeficiente de Nash-Sutcliffe para validação na maioria das simulações, com exceção da situação diária para o KNN.

Para a abordagem diária, nos modelos baseados em RNA, nota-se que o melhor valor de eficiência ocorreu na formulação 3 ($E = 0,66$), classificado como bom. Ademais, a formulação 1 apresentou um resultado próximo ($E = 0,65$), considerado satisfatório. No KNN, o melhor valor obtido ocorreu na formulação 1 ($E = 0,63$), também sendo classificado como satisfatório.

Figura 19 – Melhores índices de desempenho nas Formulações



Fonte: O autor.

Dessa maneira, a aplicação da formulação 1, tanto na RNA como no KNN, torna-se uma opção viável pois essa utiliza apenas dados de vazão como atributos de entradas, sem ser necessário fazer uso dos dados de precipitação, ou seja, é possível obter um resultado interessante a partir de uma modelagem mais simplificada.

Na situação diária também pode-se perceber que a formulação 2 foi a que apresentou os piores resultados. Ela indicou valores insatisfatórios tanto para a RNA quanto para o KNN, com $E = 0,04$ e $E = 0,39$, respectivamente. A formulação 2 é constituída basicamente por atributos de precipitação, sem fazer uso de atributos de vazão. Dessa forma, pode-se inferir que para este trabalho os atributos de precipitação talvez não possuam tanta relevância como os de vazão para a abordagem diária.

Na abordagem mensal, a formulação 3 com RNA foi a que mais se destacou dentre todas as simulações realizadas, com o resultado de 83% para o índice de eficiência para validação, sendo considerado muito bom. No KNN a formulação apresentou um resultado classificado como bom, mas distante do obtido com RNA, com um índice de eficiência de 70%.

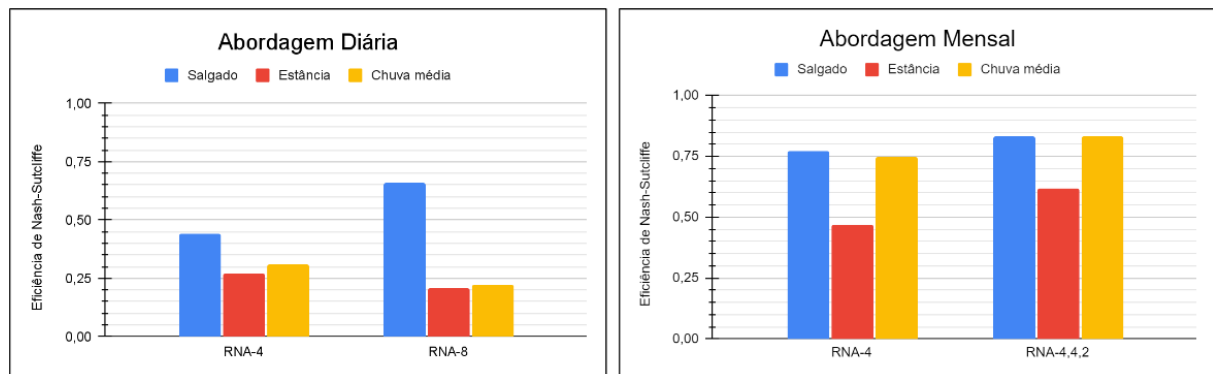
Além disso, na situação mensal a formulação 1 foi a que retratou os piores resultados. Com valores considerados insatisfatórios tanto para a RNA quanto para o KNN, com $E = 0,02$ e $E = 0,19$, respectivamente. Já a formulação 2 apresentou resultados classificados como bom ($E = 0,66$) e satisfatório ($E = 0,56$) para a RNA e KNN, respectivamente. Desse modo, é possível inferir que para este estudo os atributos de precipitação talvez tenham uma relevância superior aos atributos de vazão para a abordagem mensal.

A formulação 4 não apresentou nenhum resultado expressivo, e teve somente resultados superiores sobre a formulação 2 na abordagem diária, e sobre formulação 1 na abordagem mensal.

4.2.2 Comparação entre os postos de precipitação

Como discutido no item 3.2.1, neste estudo foi utilizado o posto fluviométrico de Estância e o pluviométrico de Salgado. Com o intuito de comparar a modelagem elaborada, testou-se novas simulações com diferentes dados de precipitação, mas ainda utilizando o posto fluviométrico

Figura 20 – Eficiências de Nash-Sutcliffe para modelagem com diferentes dados de precipitação



Fonte: O autor.

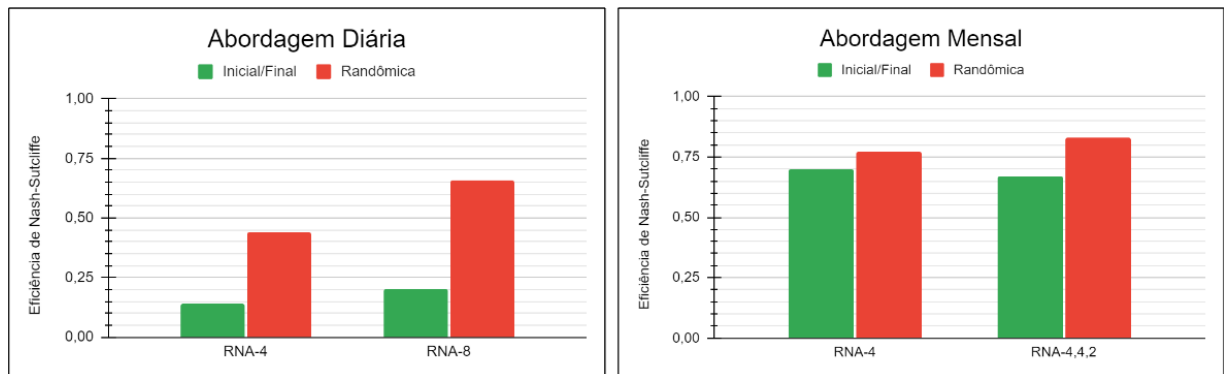
de Estância nos dados de vazão. Para isso, foi aplicada a formulação 3 com RNA, visto que esta foi a que apresentou os melhores resultados obtidos. Assim, para a análise utilizou-se a configuração automática gerada pelo Weka, e as respectivas melhores configurações obtidas neste estudo nas abordagens diária e mensal para a formulação 3. Além disso, foi aplicado o mesmo intervalo de tempo para as novas modelagens.

Na primeira simulação, para a obtenção dos dados de chuva é utilizado o posto pluviométrico de Estância. Dessa forma, a modelagem apresentou para a abordagem diária um índice de eficiência para validação de 27% para a configuração RNA-4 e 21% para a RNA-8. Ademais, para a abordagem mensal, encontraram-se $E = 0,47$ para RNA-4 e $E = 0,62$ para RNA-4,4,2. Sendo assim, a modelagem realizada utilizando o posto pluviométrico de Salgado apresentou resultados superiores, já que para a abordagem diária obteve 44% de índice de eficiência para a configuração RNA-4, e 66% para RNA-8, e para a situação mensal apresentou $E = 0,77$ e $E = 0,83$ para as configurações RNA-4 e RNA-4,4,2, respectivamente.

Na segunda simulação, para a obtenção dos dados de chuva, foi empregado a média ponderada aplicada no estudo de Lima (2022), em que o autor utilizou o método dos polígonos de Thiessen e encontrou que o posto pluviométrico de Salgado (01137001) possui influência de 77,54% da área total da bacia do rio Piauitinga, enquanto que o posto de Estância (01137017) tem influência de 22,46%. Dessa maneira, esses pesos foram aplicados nos dados de precipitação e foi encontrado que para a abordagem diária, $E = 0,31$ para a configuração RNA-4 e $E = 0,22$ para a configuração RNA-8. Já para a abordagem mensal, encontraram-se $E = 0,75$ para a configuração RNA-4 e $E = 0,83$ para a configuração RNA-4,4,2. Para a modelagem utilizando o posto pluviométrico de Salgado, atingiu-se para a abordagem diária $E = 0,44$ e $E = 0,66$ para as configurações RNA-4 e RNA-8, respectivamente, e para a abordagem mensal obteve-se $E = 0,77$ para a RNA-4 e $E = 0,83$ para a RNA-4,4,2.

Dessa maneira, pode-se notar que neste estudo a utilização da modelagem utilizando o posto pluviométrico de Salgado e o pluviométrico de Estância, é um opção viável que apresentou resultados consideráveis (Figura 20), e sua modelagem é mais simplificada que a simulação que utiliza a chuva média na bacia. A modelagem obteve resultados superiores às demais para as

Figura 21 – Influência da seleção dos dados de calibração e validação



Fonte: O autor.

configurações propostas, com exceção da situação mensal na segunda simulação, em que para a configuração RNA-4,4,2, encontraram-se valores idênticos ($E = 0,83$).

4.2.3 Influência da seleção dos dados de calibração e validação

Para observar neste estudo a influência da escolha randômica da classificação dos dados (Item 3.2.3) com 70% para calibração e 30% validação, realizou-se novas simulações em que os primeiros 70% dos dados foram classificados como calibração e os 30% finais como validação. Para isso, também foi aplicada RNA para a formulação 3, e utilizou-se a configuração automática gerada pelo Weka e as respectivas melhores configurações obtidas nas abordagens diária e mensal.

Dessa forma, para a simulação que utiliza 70% dos dados iniciais para calibração e 30% dos finais para a validação, na situação diária obteve-se $E = 0,14$ para RNA-4 e $E = 0,20$ para RNA-8. A situação mensal forneceu índices de eficiência para validação de $E = 0,70$ e $E = 0,67$ para RNA-4 e RNA-4,4,2, respectivamente. Por outro lado, na simulação que utiliza 70% dos dados de calibração e 30% para validação selecionados de maneira randômica, foram encontradas para a abordagem diária $E = 0,44$ para RNA-4 e $E = 0,66$ para RNA-8, e para abordagem mensal foi obtido $E = 0,77$ para RNA-4 e $E = 0,83$ para RNA-4,4,2.

Na Figura 21, fica evidente que os resultados obtidos utilizando a classificação randômica foram bastante superiores para as configurações analisadas, quando comparados com a classificação inicial/final.

Conclusão

O acesso de maneira adequada aos recursos hídricos é imprescindível para o desenvolvimento das regiões, entretanto o gerenciamento eficiente desses recursos pode ser uma tarefa desafiadora. A aplicação da modelagem chuva-vazão pode ser utilizada para auxiliar na gestão adequada do uso da água. Quando iniciou-se o trabalho de pesquisa, constatou-se que os modelos guiados por dados são uma opção viável para o desenvolvimento desse tipo de modelagem, com diversos exemplos na literatura que apresentaram resultados expressivos.

Neste trabalho, observou-se que é possível realizar a aplicação da modelagem chuva-vazão utilizando modelos baseados em redes neurais artificiais e no algoritmo dos k -vizinhos mais próximos na bacia hidrográfica do rio Piauitinga, em Sergipe, por meio do software Weka.

Por meio do índice de eficiência para os dados de validação, foi possível verificar o desempenho das simulações para cada configuração empregada neste estudo. Os melhores desempenhos encontrados para as abordagens diária e mensal foram atingidos a partir dos modelos baseados em RNA, e foram classificados como bom para a situação diária e muito bom para a mensal. Já nos modelos que utilizam o KNN, também foram encontrados resultados relevantes, em que obteve-se eficiência classificada como satisfatória para a abordagem diária e boa para a mensal.

Neste estudo, foram empregados quatro tipos de formulações diferentes e notou-se que a formulação 3, constituída por atributos de chuva e vazão, foi a que mais se sobressaiu, principalmente na abordagem mensal, e somente apresentou índice de eficiência inferior na situação diária utilizando KNN.

Com finalidade de comparar os resultados das simulações geradas a partir dos dados de precipitação provenientes do posto pluviométrico de Salgado, realizaram-se novas modelagens com outros dados de chuva. Dessa maneira, foram testadas duas novas simulações, uma que utilizava o pluviométrico de Estância e outra que aplicava uma média ponderada entre os postos pluviométricos de Salgado e Estância. Com isso, observou-se que a modelagem com apenas o posto pluviométrico de Salgado apresentou resultados superiores em relação a simulação com o posto de Estância, e melhores ou iguais para a simulação que utiliza a chuva média na bacia. Constatou-se, assim, que esta é uma atrativa opção, devido aos seus bons resultados e por apresentar uma modelagem mais simples quando comparada com a da chuva média.

Para investigar a influência da escolha randômica da classificação dos dados de calibração e validação neste estudo, foram feitas novas simulações utilizando os primeiros dados para calibração e os últimos para validação. Dessa forma, percebeu-se que a aplicação da classificação dos dados de forma randômica para este trabalho foi positiva, em que foram apresentados resultados bastante superiores em relação a classificação inicial/final.

Todas as simulações realizadas neste trabalho foram avaliadas a partir do coeficiente de Nash-Sutcliffe, de modo que foi necessário testar de maneira individual cada configuração em-

pregada nos modelos gerados pelo Weka, utilizando planilhas eletrônicas. Consequentemente, uma das limitações dessa pesquisa foi a quantidade de simulações que o usuário pode efetuar.

Para a contemplação dos DDMs não é necessário possuir conhecimento das características físicas da bacia, assim não é possível entender com clareza por qual motivo a formulação que utiliza somente dados de entrada de vazão, apresentou um desempenho satisfatório para abordagem diária, mas resultados insatisfatórios para a mensal. Enquanto que a formulação que utiliza apenas dados de precipitação exibiu resultados ruins para a situação diária, mas resultados bons para a mensal.

Para novas pesquisas, é recomendável a aplicação de RNAs acopladas a um algoritmo genético para realização do treinamento, comparando os resultados com os obtidos neste trabalho, e observando se ocorre melhora no desempenho da modelagem.

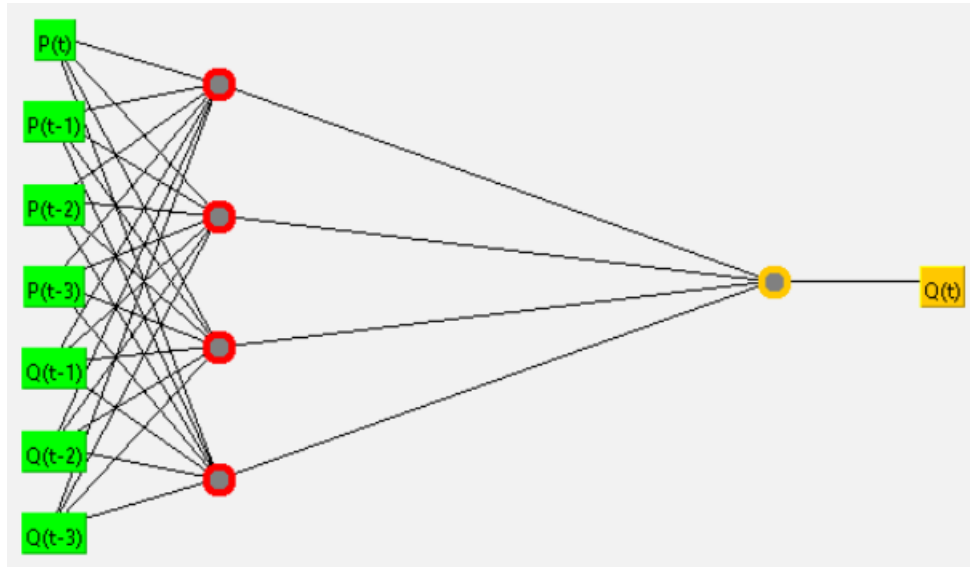
Também é recomendado o emprego de funções *kernel* na modelagem baseada em KNN, verificando se é possível atingir melhores resultados para o coeficiente de Nash-Sutcliffe.

Referências

- AKBARI, M.; OVERLOOP, P. J. v.; AFSHAR, A. Clustered k nearest neighbor algorithm for daily inflow forecasting. *Water Resources Management*, Springer, v. 25, n. 5, p. 1341–1357, 2011.
- CELESTE, A.; CHAVES, V. S. Avaliação de algoritmos de otimização e funções objetivo para calibração automática do modelo chuva-vazão tank model. *Ciência e Natura*, v. 36, n. 3, p. 527–537, 2014.
- CELESTE, A. B.; MENEZES, T.; FONTES, V. Modelagem do processo chuva-vazão pelo tank model e por redes neurais artificiais na bacia do rio Japarutuba, Sergipe. *Scientia Plena*, v. 10, n. 7, 2014.
- FARIAS, C. A. S.; SANTOS, C. A. G.; CELESTE, A. B. Rede neural com atraso de tempo aplicada à previsão mensal de vazões na bacia hidrográfica do rio Piancó. In: Associação Brasileira de Recursos Hídricos (ABRH). *Proc. X Simpósio de Recursos Hídricos do Nordeste*. Fortaleza, Brazil, 2010.
- FRANK, E.; HALL, M. A.; WITTEN, I. H. *The WEKA Workbench. Online Appendix for "Data Mining: Practical Machine Learning Tools and Techniques"*. Fourth edition. [S.l.]: Morgan Kaufmann, 2016.
- LIMA, R. G. S. *Modelo chuva-vazão baseado na aplicação do método dos resíduos ponderados ao balanço hídrico*. São Cristóvão – SE, 2022. 33 f. Trabalho de Conclusão de Curso (TCC) – Departamento de Engenharia Civil. Universidade Federal de Sergipe.
- MACHADO, F. W. *Modelagem chuva-vazão mensal utilizando redes neurais artificiais*. Dissertação (Mestrado) – Programa de Pós-Graduação em Engenharia de Recursos Hídricos e Ambiental, Universidade Federal do Paraná, 2005.
- SCHEIDT, F. A.; BRUNETTO, M. A. de C. Modelagem chuva-vazão utilizando redes neurais artificiais e algoritmos genéticos. In: *Congresso da Sociedade Brasileira de Computação*. [S.l.: s.n.], 2011. v. 31.
- SEMARH-SE. *Atlas digital sobre recursos hídricos de Sergipe*. 2012. CD-ROM.
- SETTI, A. A. et al. *Introdução ao Gerenciamento de Recursos Hídricos*. 2º. ed. [S.l.]: Brasília: Agência Nacional de Energia Elétrica, Superintendência de Estudos e Informações Hidrológicas, 2001.
- SOLOMATINE, D. P.; MASKEY, M.; SHRESTHA, D. L. Instance-based learning compared to other data-driven methods in hydrological forecasting. *Hydrological Processes: An International Journal*, Wiley Online Library, v. 22, n. 2, p. 275–287, 2008.
- SOUSA, F.; SOUSA, W. Rede neural aplicada ao processo chuva-vazão da bacia hidrográfica de Caraúbas-PB. In: *Congresso Brasileiro de Meteorologia*. [S.l.: s.n.], 2006. v. 14.
- VIEGAS FILHO, J. S. et al. MODHAC – modelo hidrológico auto-calibrável, versão 98 para Windows. In: *Anais do Congresso Brasileiro de Engenharia Agrícola*. Pelotas: [s.n.], 1999.

Apêndices

APÊNDICE A – Exemplo de visualização gerada pelo Weka para as RNAs



APÊNDICE B – Exemplo de pesos e funções de ativação utilizados durante o processo de treinamento de uma RNA no Weka

```

=== Classifier model (full training set) ===

Linear Node 0
  Inputs      Weights
  Threshold   0.5068338725109756
  Node 1      -1.6734425581665322
  Node 2       0.8243740561809545
  Node 3       0.8027649951415559
  Node 4       0.8050832468719418

Sigmoid Node 1
  Inputs      Weights
  Threshold   2.7982314602386262
  Attrib P(t) -2.705819499272237
  Attrib P(t-1) 1.1640985746026156
  Attrib P(t-2) -1.3259940501695155
  Attrib P(t-3) -0.1173618323676015
  Attrib Q(t-1) -1.2466903938929497
  Attrib Q(t-2) -0.4791984453364398
  Attrib Q(t-3) 2.832809060186523

Sigmoid Node 2
  Inputs      Weights
  Threshold   -0.6272484798365587
  Attrib P(t) 0.9170593390234021
  Attrib P(t-1) 0.4262370893703075
  Attrib P(t-2) -0.7791233475396021
  Attrib P(t-3) -0.1992746984860193
  Attrib Q(t-1) 0.4299954495202029
  Attrib Q(t-2) 0.443805961913883
  Attrib Q(t-3) 0.8350593959084387

Sigmoid Node 3
  Inputs      Weights
  Threshold   -0.8213908572642292
  Attrib P(t) 0.21414087675997445
  Attrib P(t-1) 0.2915525535330575
  Attrib P(t-2) 0.29235431952093094
  Attrib P(t-3) 0.22878373452126982
  Attrib Q(t-1) 0.21321575729281542
  Attrib Q(t-2) 0.047065370999001654
  Attrib Q(t-3) 0.36526846142170427

Sigmoid Node 4
  Inputs      Weights
  Threshold   -0.615123051057784
  Attrib P(t) 1.0657298575810983
  Attrib P(t-1) 0.3185343928239416
  Attrib P(t-2) -0.8093293556774074
  Attrib P(t-3) -0.20127200120937977
  Attrib Q(t-1) 0.5288813432201219
  Attrib Q(t-2) 0.5188496864046925
  Attrib Q(t-3) 0.7424290260460276

```

APÊNDICE C – Eficiências de Nash-Sutcliffe nas Formulações 1 e 2 para a abordagem diária com RNA

Configuração	Formulação 1			Formulação 2		
	Completo	Calibração	Validação	Completo	Calibração	Validação
RNA-3	0,53	0,52	0,54	0,18	0,30	-0,22
RNA-3,1	0,56	0,54	0,62	0,08	0,25	-0,43
RNA-3,3	0,57	0,55	0,64	0,12	0,22	-0,17
RNA-3,1,1	-0,01	-0,01	-0,01	0,02	0,05	-0,07
RNA-3,3,1	-0,01	-0,01	-0,01	-0,01	-0,01	-0,01
RNA-3,3,3	0,56	0,54	0,62	0,07	0,14	-0,14
RNA-4	0,52	0,52	0,52	0,21	0,33	-0,17
RNA-4,2	0,57	0,54	0,63	0,11	0,19	-0,12
RNA-4,4	0,57	0,55	0,64	0,13	0,21	-0,14
RNA-4,2,1	-0,01	-0,01	-0,01	0,00	0,00	0,00
RNA-4,4,2	0,56	0,54	0,62	0,06	0,11	-0,11
RNA-4,4,4	0,56	0,54	0,63	0,10	0,17	-0,10
RNA-6	0,51	0,53	0,45	0,42	0,54	0,04
RNA-6,3	0,57	0,55	0,65	0,13	0,20	-0,1
RNA-6,6	0,57	0,55	0,65	0,12	0,18	-0,06
RNA-6,3,3	0,56	0,54	0,62	0,09	0,15	-0,10
RNA-6,6,3	0,56	0,54	0,63	0,11	0,17	-0,08
RNA-6,6,6	0,57	0,55	0,63	0,11	0,21	-0,18

APÊNDICE D – Eficiências de Nash-Sutcliffe nas Formulações 1 e 2 para a abordagem mensal com RNA

Configuração	Formulação 1			Formulação 2		
	Completo	Calibração	Validação	Completo	Calibração	Validação
RNA-3	-0,04	-0,05	-0,03	0,77	0,81	0,61
RNA-3,1	0,00	0,02	-0,03	0,74	0,78	0,59
RNA-3,3	-0,01	0,01	-0,04	0,74	0,78	0,60
RNA-3,1,1	-0,48	-0,52	-0,41	0,73	0,76	0,61
RNA-3,3,1	-0,48	-0,52	-0,41	-0,05	-0,05	-0,06
RNA-3,3,3	-0,48	-0,52	-0,41	0,75	0,78	0,62
RNA-4	-0,05	-0,06	-0,03	0,76	0,81	0,56
RNA-4,2	-0,01	0,01	-0,04	0,76	0,79	0,63
RNA-4,4	-0,01	0,00	-0,04	0,76	0,79	0,63
RNA-4,2,1	-0,48	-0,52	-0,41	-0,05	-0,05	-0,06
RNA-4,4,2	-0,48	-0,52	-0,41	0,76	0,78	0,65
RNA-4,4,4	0,00	0,02	-0,03	0,76	0,78	0,66
RNA-6	-0,05	-0,06	-0,04	0,76	0,81	0,58
RNA-6,3	0,00	0,01	-0,03	0,76	0,79	0,64
RNA-6,6	-0,01	0,01	-0,04	0,77	0,80	0,64
RNA-6,3,3	-0,48	-0,52	-0,41	0,76	0,79	0,66
RNA-6,6,3	0,05	0,07	0,02	0,76	0,78	0,66
RNA-6,6,6	0,02	0,04	-0,01	0,75	0,78	0,64

APÊNDICE E – Eficiências de Nash-Sutcliffe nas Formulações 3 e 4 para a abordagem diária com RNA

Configuração	Formulação 3			Formulação 4		
	Completo	Calibração	Validação	Completo	Calibração	Validação
RNA-4	0,56	0,67	0,44	0,57	0,65	0,48
RNA-4,2	0,61	0,60	0,63	0,59	0,61	0,58
RNA-4,4	0,62	0,60	0,63	0,59	0,61	0,57
RNA-4,2,1	0,56	0,53	0,60	0,56	0,57	0,56
RNA-4,4,2	0,59	0,56	0,62	0,54	0,56	0,51
RNA-4,4,4	0,62	0,60	0,64	0,59	0,61	0,57
RNA-6	0,62	0,63	0,61	0,55	0,62	0,46
RNA-6,3	0,60	0,58	0,63	0,59	0,60	0,58
RNA-6,6	0,61	0,59	0,63	0,59	0,60	0,57
RNA-6,3,3	0,62	0,59	0,64	0,59	0,59	0,58
RNA-6,6,3	0,61	0,59	0,64	0,59	0,60	0,58
RNA-6,6,6	0,59	0,56	0,62	0,59	0,60	0,58
RNA-8	0,64	0,62	0,66	0,48	0,65	0,30
RNA-8,4	0,60	0,58	0,62	0,59	0,60	0,58
RNA-8,8	0,61	0,59	0,63	0,59	0,60	0,57
RNA-8,4,4	0,60	0,57	0,63	0,59	0,59	0,58
RNA-8,8,4	0,61	0,58	0,63	0,58	0,59	0,58
RNA-8,8,8	0,60	0,58	0,63	0,59	0,60	0,58

APÊNDICE F – Eficiências de Nash-Sutcliffe nas Formulações 3 e 4 para a abordagem mensal com RNA

Configuração	Formulação 3			Formulação 4		
	Completo	Calibração	Validação	Completo	Calibração	Validação
RNA-4	0,82	0,83	0,77	-0,49	0,52	-4,21
RNA-4,2	0,81	0,82	0,78	0,38	0,55	-0,22
RNA-4,4	0,81	0,82	0,81	0,34	0,45	-0,06
RNA-4,2,1	0,00	0,00	-0,01	-0,05	-0,05	-0,07
RNA-4,4,2	0,82	0,82	0,83	0,26	0,27	0,24
RNA-4,4,4	0,82	0,82	0,81	0,31	0,31	0,29
RNA-6	0,72	0,76	0,55	-0,48	0,52	-4,15
RNA-6,3	0,82	0,82	0,82	0,38	0,49	-0,03
RNA-6,6	0,82	0,82	0,81	0,36	0,51	-0,21
RNA-6,3,3	0,82	0,82	0,83	0,32	0,33	0,29
RNA-6,6,3	0,79	0,78	0,81	0,29	0,28	0,29
RNA-6,6,6	0,82	0,82	0,81	0,27	0,27	0,27
RNA-8	0,72	0,76	0,50	-0,44	0,53	-4,00
RNA-8,4	0,81	0,81	0,79	0,32	0,35	0,22
RNA-8,8	0,81	0,81	0,80	0,31	0,34	0,21
RNA-8,4,4	0,82	0,82	0,82	0,32	0,35	0,23
RNA-8,8,4	0,82	0,82	0,83	0,29	0,29	0,29
RNA-8,8,8	0,82	0,82	0,81	0,29	0,30	0,26

APÊNDICE G – Eficiências de Nash-Sutcliffe nas Formulações 1 e 2 para o KNN

Abordagem Diária						
Configuração	Formulação 1			Formulação 2		
	Completo	Calibração	Validação	Completo	Calibração	Validação
1–NN	0,83	1,00	0,32	0,62	0,99	-0,55
3–NN	0,63	0,69	0,45	0,54	0,60	0,34
5–NN	0,61	0,62	0,59	0,47	0,50	0,36
7–NN	0,58	0,57	0,60	0,43	0,44	0,39
10–NN	0,56	0,55	0,61	0,41	0,41	0,39
13–NN	0,56	0,55	0,59	0,37	0,37	0,38
15–NN	0,57	0,56	0,62	0,37	0,37	0,37
17–NN	0,56	0,54	0,62	0,36	0,35	0,37
20–NN	0,55	0,53	0,63	0,34	0,34	0,36

Abordagem Mensal						
Configuração	Formulação 1			Formulação 2		
	Completo	Calibração	Validação	Completo	Calibração	Validação
1–NN	0,64	1,00	0,09	0,89	1,00	0,46
3–NN	0,38	0,57	0,11	0,38	0,39	0,35
5–NN	0,31	0,42	0,13	0,68	0,71	0,56
7–NN	0,32	0,40	0,19	0,65	0,67	0,54
10–NN	0,30	0,38	0,18	0,62	0,64	0,54
13–NN	0,30	0,39	0,17	0,59	0,61	0,53
15–NN	0,29	0,38	0,17	0,58	0,59	0,54
17–NN	0,30	0,37	0,19	0,57	0,58	0,52
20–NN	0,28	0,35	0,17	0,56	0,56	0,53

APÊNDICE H – Eficiências de Nash-Sutcliffe nas Formulações 3 e 4 para o KNN

Abordagem Diária						
Configuração	Formulação 3			Formulação 4		
	Completo	Calibração	Validação	Completo	Calibração	Validação
1–NN	0,72	1,00	0,42	0,67	1,00	0,32
3–NN	0,59	0,71	0,45	0,55	0,71	0,39
5–NN	0,56	0,67	0,43	0,50	0,63	0,37
7–NN	0,53	0,61	0,43	0,50	0,58	0,41
10–NN	0,51	0,57	0,43	0,46	0,53	0,39
13–NN	0,47	0,54	0,39	0,44	0,50	0,37
15–NN	0,45	0,51	0,38	0,42	0,48	0,36
17–NN	0,47	0,54	0,39	0,40	0,46	0,34
20–NN	0,43	0,48	0,37	0,40	0,45	0,35

Abordagem Mensal						
Configuração	Formulação 3			Formulação 4		
	Completo	Calibração	Validação	Completo	Calibração	Validação
1–NN	0,92	1,00	0,53	0,47	1,00	-1,49
3–NN	0,76	0,80	0,59	0,49	0,61	0,07
5–NN	0,71	0,72	0,63	0,42	0,47	0,23
7–NN	0,67	0,67	0,66	0,36	0,36	0,33
10–NN	0,66	0,65	0,69	0,33	0,33	0,36
13–NN	0,64	0,62	0,70	0,33	0,32	0,36
15–NN	0,63	0,62	0,69	0,32	0,31	0,35
17–NN	0,62	0,60	0,69	0,32	0,31	0,37
20–NN	0,59	0,58	0,67	0,33	0,32	0,38