Universidade Estadual de Campinas Faculdade de Tecnologia

GUILHERME MASAO TSUYUKUBO

UTILIZAÇÃO DE APRENDIZADO DE MÁQUINA PARA A PREVISÃO DA VAZÃO DE RIOS DE CAMPINAS

LIMEIRA
2021

Universidade Estadual de Campinas Faculdade de Tecnologia

GUILHERME MASAO TSUYUKUBO

UTILIZAÇÃO DE APRENDIZADO DE MÁQUINA PARA A PREVISÃO DA VAZÃO DE RIOS DE CAMPINAS

Trabalho de Conclusão do Curso apresentado como requisito parcial para a obtenção do título de Bacharel em Sistemas de Informação à Faculdade de Tecnologia da Universidade Estadual de Campinas.

Orientadora: Prof.ª Dr.ª Ana Estela Antunes da Silva

LIMEIRA
2021

Ficha catalográfica
Universidade Estadual de Campinas
Biblioteca da Faculdade de Tecnologia
Luiz Felipe Galeffi - CRB 8/10385

Tsuyukubo, Guilherme Masao, 1998-
T789a Utilização de aprendizado de máquina para a previsão da vazão de rios de Campinas / Guilherme Masao Tsuyukubo. – Limeira, SP : [s.n.], 2021.

Orientador: Ana Estela Antunes da Silva.
Trabalho de Conclusão de Curso (graduação) – Universidade Estadual de Campinas, Faculdade de Tecnologia.

1. Mineração de dados (Computação). 2. Aprendizado de máquina. 3. Medidores de fluxo. I. Silva, Ana Estela Antunes da,1965-. II. Universidade Estadual de Campinas. Faculdade de Tecnologia. III. Título.

Informações adicionais, complementares

Título em outro idioma: Use of machine learning to forecast Campinas river flow
Palavras-chave em inglês:
Data mining
Machine learning
Flow meters
Titulação: Bacharel
Banca examinadora:
Ana Estela Antunes da Silva [Orientador]
Guilherme Palermo Coelho
Lubienska Cristina Lucas Jaquiê Ribeiro
Data de entrega do trabalho definitivo: 21-12-2021

FOLHA DE APROVAÇÃO

Abaixo se apresentam os membros da comissão julgadora da sessão pública de defesa de dissertação para o Título de Bacharel em Sistemas de Informação na área de concentração, a que se submeteu o aluno Guilherme Masao Tsuyukubo, em 21 de dezembro de 2021 na Faculdade de Tecnologia – FT/UNICAMP, em Limeira/SP.

Prof. Dra. Ana Estela de Antunes da Silva
Presidente da Comissão Julgadora

Prof. Guilherme Palermo Coelho
FT/Unicamp

Prof. Dra. Lubienska Cristina Lucas Jaquiê Ribeiro
FT/Unicamp

Ata da monografia, assinada pelos membros da Comissão Examinadora, consta no SIGA/Trabalho de Conclusão de Curso/Monografia e na Secretaria de Graduação da FT – Unicamp.

RESUMO

A ocorrência de enchentes traz transtornos para as comunidades que vivem em torno de regiões com bacias hídricas próximas. A fim de evitar esse problema e, consequentemente, melhorar o sistema de gerenciamento, a predição da vazão de rios torna-se fundamental. O presente estudo faz a utilização das Redes Perceptron Multi-Camadas (MLP) com o objetivo de obter um modelo de previsão de vazão. Para alcançar essa meta, foi utilizada uma base de dados de uma bacia hídrica da região de Campinas, localizada no interior de São Paulo. Os resultados obtidos demonstram a eficiência do modelo em relação à vazão hídrica dessa bacia hidrográfica. A base inicial foi dividida em diária e mensal e, foi utilizada a técnica SMOTE para aumentar o volume de dados das bases. Os modelos, no final, apresentaram acurácia média de 95% no caso da base mensal e 99% em relação à base diária.

Palavras-chave: Mineração de dados, Aprendizado de Máquina, Vazão dos Rios.

ABSTRACT

The occurrence of floods brings inconvenience to the communities that live around regions with nearby water basins. To avoid this problem and, consequently, improve the management system, the prediction of river flow becomes fundamental. The present study makes use of Multi-layer Perceptron Networks (MLP) with the goal of obtaining a model for flow prediction. To achieve this goal, a data base from a water basin in the region of Campinas, located in the interior of São Paulo State, was used. The results obtained demonstrate the efficiency of the model in relation to the hydrological flow of this hydrographic basin. The initial database was divided into daily and monthly databases, and the SMOTE technique was applied to increase the database volume. The models average accuracy was 95% in the monthly basis case and 99% in relation to the daily basis.

Keywords: Data mining, Machine Learning, River flow.

LISTA DE ILUSTRAÇÕES

LISTA DE TABELAS

LISTA DE ABREVIATURA E SIGLAS

MMA	Ministério do Meio Ambiente
ANA	Agência Nacional de Águas e Saneamento Básico
TI	Tecnologia da Informação
RN	Redes Neurais
MLP	Perceptron Multicamadas
RNCs	Redes Neurais Convolucionais
RNRs	Redes Neurais Recorrentes
LSTM	Long Short-Term Memory
LMS	Least-Mean Square
ANN	Autoencoder Neural Network
SVM	Máquinas de Vetor de Suporte
ELM	Máquinas de Aprendizado Extremo
ARIMA	Integrated Autoregressive Moving Around Model
PCA	Principal Component Analysis
SOM	Self-Organizing Maps
RF	Random Forest

SUMÁRIO

1. INTRODUÇÃO

No Brasil é cada vez mais comum ouvir relatos de problemas envolvendo enchentes. Isso porque, em meses em que há uma alta carga pluviométrica, há uma sobrecarga dos sistemas de escoamento. Isso pode ocorrer por diversos fatores sendo um deles o processo de urbanização acelerado e mal planejado. Além disso, há alterações nas margens de rios que, por natureza, possuem uma área de escape para uma possível demanda exacerbada. Com isso, a vazão ultrapassa a capacidade de escoamento e ocasiona um aumento na quantidade de alagamentos (POLI, 2013).

Além do problema das inundações, a alteração na vazão de um rio que possui conexão com o oceano pode resultar em uma alteração na densidade da circulação de uma região, efeito ocasionado pelo processo de evaporação da superfície marítima. Um outro aspecto afetado é a bioquímica das águas marítimas, isso porque, um aumento na vazão resulta em um desequilíbrio na foz, impactando em todo o ecossistema do local (ZENG et al., 2012). Portanto, uma simples variação pode impactar tanto no âmbito social quanto natural.

Por conta disso, a possibilidade de prever de forma precisa a vazão dos rios das regiões torna-se um aliado útil. Isso porque, a partir desse estudo há a possibilidade da otimização dos sistemas de armazenamento de água, a prevenção a eventos naturais como inundações e enchentes e uma melhora na eficiência de geração elétrica, além de outros muitos benefícios. A interação entre os humanos e os sistemas naturais nas diversas esferas é um ponto que pode dificultar o aperfeiçoamento do sistema. A análise de dados pode desempenhar um papel fundamental nesse processo de planejamento da gestão da vazão dos rios (TORO et al., 2013).

Uma base de dados bem organizada pode permitir previsões de ocorrências de eventos. Além disso, é possível a extração de padrões ocultos relacionados a eventos passados que, posteriormente, podem auxiliar no processo de criação de um sistema de apoio à decisão, com base em métodos de mineração de dados.

Um problema encontrado durante esse processo de criação do conhecimento é o pré-processamento. Para a execução de um método de mineração de dados algumas condições devem ser satisfeitas sendo algumas delas: limpeza de dados

inconsistentes e ausência de registros. Essas anomalias devem ser corrigidas, anteriormente, durante o processo de pré-processamento, para que os dados corretos possam ser utilizados posteriormente em um modelo (GILBERT; SÀNCHEZ-MARRÈ; IZQUIERDO, 2016).

A utilização de dados na predição da vazão pode, portanto, auxiliar em diversos âmbitos como, por exemplo, na estimativa do processo de erosão das margens dos rios, na mitigação de inundações ou na definição da capacidade destinada a um reservatório de emergência para possíveis crises. Além disso, um estudo desse problema pode auxiliar no trabalho de hidrólogos, especialistas em sistemas de gestão da água ou especialistas em predição de enchentes, por exemplo (YASEEN et al., 2020).

2. OBJETIVO GERAL

Utilizar métodos de Aprendizado de Máquina em bases de dados da área ambiental da região de Campinas que possui como principal atributo a vazão de rios. Além disso, criar dois modelos preditivos para auxiliar no processo de planejamento da vazão. Por fim, realizar a comparação entre os modelos criados a partir da base diária e da base mensal.

2.1 OBJETIVOS ESPECÍFICOS

- Criar a base de dados mensal;
- Criar a base de dados diária;
- Pré-processar as bases;
- Avaliar resultados da execução dos algoritmos.

3. REFERENCIAL TEÓRICO

Toda a fundamentação teórica desse trabalho será apresentada nesse capítulo. Serão discutidos tópicos como a vazão de rios, predição em aprendizado de máquinas e outros pontos.

3.1 VAZÃO DE RIOS

Entender a vazão de um rio é fundamental para evitar diversos problemas que podem ocorrer pela sobrecarga desse sistema como, por exemplo, inundações. Além disso, essa taxa está diretamente ligada à composição química da água e, assim, qualquer alteração pode resultar em uma variação na quantidade de oxigênio dissolvido, no pH ou na temperatura da água. Um outro ponto é que pode ocorrer uma modificação na absorção de determinados minerais como Nitrato e Fósforo. Por fim, a vazão pode impactar, também, na quantidade de nutrientes contidos na água visto que eles, naturalmente, tendem a ir para o fundo do rio (AKINER; AKKOYUNLU, 2012).

Um fator que impacta de forma significativa na vazão são as alterações humanas realizadas no ambiente. O consumo desenfreado de água aliado ao despejo de esgotos nos rios causam uma degradação ambiental, enquanto atividades agrícolas, irrigação e ocupação urbana, promovem alterações hídricas no local afetado. Por isso é importante ter a capacidade de entender o comportamento da vazão para realizar um planejamento mais consciente e com uma maior precisão (SOUZA; SOUZA; CARDOSO, 2017).

Apesar da grande importância de um estudo mais aprofundado na predição da vazão em países em desenvolvimento, como é o caso do Brasil, há uma grande dificuldade na realização dessas pesquisas por conta de restrições financeiras e técnicas o que prejudica no processo de coleta e distribuição dos dados. Além disso, majoritariamente, essas bases são capturadas com registros esparsos, escassos ou de baixa qualidade. Por conta disso, há uma grande dificuldade no desenvolvimento de trabalhos em relação a esses temas nesses países (BOU-FAKHREDDINE et al., 2018).

Não existe oficialmente uma categorização para os estudos envolvendo recursos hídricos, entretanto, os pesquisadores citam três possíveis grupos: a curto, médio e longo prazo. As predições a curto prazo, podem ser separadas em 3

subgrupos: por hora, por dia e por semanas. Para que esteja enquadrada no primeiro subgrupo, a previsão deve estar em um intervalo de no máximo 48 horas. Já, para estar contido no estudo diário, o prazo é de 14 dias. Por fim, a predição por semanas possui o limite de 26 semanas. Esse tipo de estudo é recomendado para se fazer o planejamento operacional (TIWARI; ADAMOWSKI, 2013; SEO; KWON; CHOI, 2018).

São considerados de médio prazo, os estudos que realizam a predição mensal dentro de um prazo de até 24 meses. Esse tipo de metodologia é a menos comum de ser utilizada. Já os estudos a longo prazo levam em consideração predições anuais e de décadas. Essa estrutura visa auxiliar na solução de problemas de maior escala em relação a água como, por exemplo, as variações climáticas, alterações econômicas e demográficas, pois, podem auxiliar no desenvolvimento e na melhora na infraestrutura de todo o sistema de planejamento. Além disso, essas pesquisam podem ser utilizadas para a predição da possibilidade de escassez desse recurso a fim de conscientizar e pensar em formas de preservá-lo (TIWARI; ADAMOWSKI, 2013; SEO; KWON; CHOI, 2018).

3.2 PREDIÇÃO EM APRENDIZADO DE MÁQUINA

O estudo da predição da vazão é fundamental para auxiliar na tomada de decisão para a realização de ações que mantenham todo o balanço do ecossistema envolvido na região em que há a passagem de uma bacia hidrográfica. Isso porque, uma pequena alteração nesse equilíbrio pode resultar em diversos fatores que impactam o entorno da região. A falta de pesquisa nessa área pode resultar em graves impactos ao local como, por exemplo, a ocorrência de graves inundações e enchentes. Além disso, entender as oscilações nessa taxa possibilita prever possíveis problemas que possam ocorrer na composição química da água. Portanto, uma pesquisa para predizer a vazão é benéfica tanto para a sociedade quanto para o meio ambiente (AKINER; AKKOYUNLU, 2012).

A previsão da distribuição hídrica é essencial para o funcionamento de todo o sistema de abastecimento de água. Isso porque, a partir dessa predição, todo o processo de planejamento é beneficiado, possibilitando assim, a entrega do recurso com uma melhor qualidade para os consumidores finais. Um outro ponto é que a água e o consumo energético são diretamente proporcionais, ou seja, a perda de um desses bens implica em prejuízo do outro. Logo, a estimativa desse recurso é benéfica para

o fator ambiental e para os seus usuários (KAMIŃSKI; KAMIŃSKI; MIZERSKI, 2017; DE MARCHIS et al., 2016).

A mineração de dados é um grande aliado nesse processo já que essa técnica possibilita a extração de informações relevantes de grandes bases de dados e, com isso, é possível realizar previsões futuras, tomar decisões com o conhecimento adquirido e entender comportamentos (THAMILSELVAN; SATHIASEELAN, 2015).

Uma ferramenta da mineração de dados que auxilia nesse processo de previsão da demanda da água são as Redes Neurais (RN). Elas são modelos de dados estatísticos não lineares baseados no funcionamento dos neurônios biológicos. Em suma, as RN são unidades de processamento simples interligadas que recebem várias entradas e retornam, na maioria das vezes, uma saída. Para isso, a informação é transmitida dentro das camadas da rede e, durante esse processo, as entradas são ponderadas e adicionadas até que, ao final, o resultado obtido passa por uma função de ativação que libera uma saída (SHARMA; RAI; DEV, 2012).

O processo de aprendizado das RN pode ser classificado em três diferentes categorias: aprendizado supervisionado, não supervisionado e por reforço. O primeiro, assim como sugere o nome, tem o auxílio de um instrutor que realiza a classificação dos dados de treino, que são valores utilizados para o processo de criação de um modelo de RN. Ou seja, você possui uma base com os resultados já definidos e, a partir desses dados, a RN aprende a classificar as saídas. Um exemplo que pode ser classificado dessa maneira é a MultiLayer Perceptron (MLP) (SATHYA; ABRAHAM, 2013).

Em contrapartida, uma RN não supervisionada é aquela em que os comportamentos são obtidos de maneira heurística. Ou seja, essa metodologia refere-se à habilidade da rede em aprender a procurar por padrões ocultos a fim de agrupar dados que ainda não foram categorizados. Por fim, o aprendizado por reforço tem o seu processo definido por iterações de tentativas e erros com o ambiente ao realizar atribuições de recompensa e penalidade a fim de se atingir um objetivo final (SATHYA; ABRAHAM, 2013).

As RN podem, também, ser classificadas em algumas categorias. Uma destas são as Redes Neurais Convolucionais (RNCs), que são redes com múltiplas camadas projetadas principalmente para o auxílio na classificação de imagens 2D. Durante a criação desse modelo, primeiramente, há a captação de uma imagem na camada de

entrada. Em seguida, há a atribuição de pesos e vieses à aspectos e objetos da imagem a fim de conseguir discernir uma das outras (ZHANG et al., 2015).

Outra categoria de RN são as Redes Neurais Recorrentes (RNRs), que diferente das demais não consideram as amostras como independentes, ou seja, atribui importância para as altas correlações dentro do material estudado, como, por exemplo, os frames de um vídeo ou palavras em uma frase. Um exemplo dessa metodologia é a Long Short-Term Memory (LSTM) (WANG; WANG; LUI, 2018).

Um terceiro tipo de RN são as Redes Neurais Feed Forward. Nessa metodologia, as informações são transmitidas através das camadas da rede e, durante esse processo, pesos são ajustados a fim de chegar ao resultado. Um método utilizado durante esse processo de ajustes é conhecido como least-mean-square (LMS) (MINAMOTO et al, 2016).

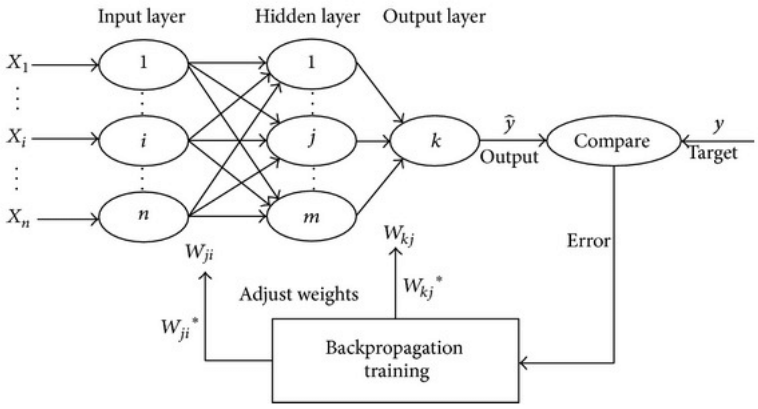
3.3 MULTILAYER PERCEPTRON

A MultiLayer Perceptron (MLP) é uma rede neural que tem várias entradas reais que, durante a passagem pelas camadas internas, passam por um processo de combinação linear com base no peso das entradas e, por fim, a saída é liberada por uma função de ativação não linear (MEDEIROS et al., 2016).

O processamento da MLP ocorre em duas fases. A primeira, conhecida como propagação, é a etapa em que as informações fluem de camada em camada dentro da rede até chegarem na saída. Além disso, durante essa etapa, os pesos são fixos. A outra, chamada de adaptação, é o processo em que a rede realiza o ajuste dos pesos (AMBROSIO, 2019).

Para a realização dos ajustes, o método mais utilizado na MLP é o de *backpropagation*. Esse algoritmo é dividido em duas etapas: inicialmente, há a realização de um processo para frente em que um vetor de entrada é aplicado e propagado em todos os nós dentro de todas as camadas até o final da rede. Com isso, é realizada a comparação entre a resposta obtida e a resposta esperada e, há o cálculo do erro. Esse valor é utilizado para realizar ajustes nos pesos sinápticos da rede. Os pesos sinápticos são aplicados sempre em que há a passagem dos dados de uma camada para a outra da rede. Por fim, há a passagem do sinal do erro da saída para a entrada. Esse processo é realizado até se atingir o critério de parada (MANZAN, 2016). Uma ilustração do método é o representado na Figura 1.

Figura 1 – Método de *backpropagation*



Fonte: (GHORBANI; BARARI; HOSEINI, 2018)

3.4 TRATAMENTO DE DADOS

Para possibilitar a utilização de dados de melhor qualidade como de uma MLP, o processo de tratamento de dados é fundamental. Isso porque, nessa etapa há a preparação dos dados para que a análise e a execução posterior das técnicas de aprendizado de máquina possam ser realizadas de maneira eficaz. Essa área possui diversos tipos de técnicas associadas a ela, dentre as quais podem ser citadas, por exemplo, a discretização, a limpeza dos dados e a normalização (DE CASTRO; FERRARI, 2016). Um outro método comumente utilizado é o do *oversampling*.

A limpeza de dados refere-se ao processo de extração dos valores que de alguma forma podem interferir no processo de criação dos modelos de dados. Dentre algumas técnicas de limpeza de dados que visam evitar a utilização de valores com baixa qualidade podem ser citados: correção de dados ausentes, suavização de ruídos e identificação de valores discrepantes (DE CASTRO; FERRARI, 2016).

Um ponto importante para este estudo é a imputação de dados ausentes. Em relação a esse processo, algumas das técnicas são: ignorar o objeto, imputar manualmente os valores ausentes e imputação do tipo *hot-deck*. O primeiro método consiste em remover todos os objetos que possuam algum valor ausente na base. Já, a imputação manual é a escolha e a inserção de um valor empírico para substituir os valores ausentes (DE CASTRO; FERRARI, 2016).

Um outro ponto relevante para este estudo é o processo de discretização. A discretização é utilizada para realizar a conversão de atributos numéricos em categóricos (DE CASTRO; FERRARI, 2016).

Existem diversas técnicas de discretização de dados sendo a mais utilizada a de encaixotamento. Além do processo de discretização, esse método é utilizado no tratamento de dados ruidosos. O encaixotamento pode ser subdividido em dois grupos: de mesma largura e o de mesma frequência (DE CASTRO; FERRARI, 2016).

O primeiro consiste em dividir os dados do atributo em caixas com o mesmo intervalo, ou seja, em agrupamentos em que a métrica utilizada para realizar essa separação seja a mesma para todos os conjuntos de dados. Por exemplo, na divisão de registros pertencentes ao intervalo de 3 a 23, as caixas podem ser subdivididas em caixa 1 sendo o conjunto de números entre 3 e 13 e a outra contendo os dados entre 13 e 23. Em ambos os casos, mantêm-se a diferença de 10, preservando assim, o tamanho da caixa (DE CASTRO; FERRARI, 2016).

Já o intuito do encaixotamento de mesma frequência é o de distribuir os dados de forma equânime entre os conjuntos, ou seja, com o mesmo número de componentes em todos os grupos. Por exemplo, a separação de seis dados distintos em duas caixas resultaria em dois conjuntos com três itens cada, mantendo, assim, a quantidade de itens contidos em cada agrupamento (DE CASTRO; FERRARI, 2016).

Algumas outras técnicas de discretização são: por histograma e por algoritmos de agrupamento. A primeira é similar à metodologia de encaixotamento com a diferença de que, nesse caso, as faixas do histograma são utilizadas para fazer a divisão dos dados nos grupos. Já na técnica de discretização por algoritmos de agrupamento, os valores são divididos em grupos a partir de um critério preestabelecido e, para fazer a divisão, cada um desses grupos é representado por um protótipo. Essa metodologia pode ser realizada por diferentes tipos de algoritmos de agrupamento (DE CASTRO; FERRARI, 2016).

Após o processo de discretização e de limpeza, a base ainda pode permanecer desbalanceada com algumas classes possuindo uma menor quantidade de objetos em relação às outras. A fim de minimizar esse problema, as técnicas de *oversampling* podem ser utilizadas. O método mais conhecido é o do SMOTE que consiste em replicar os elementos da classe minoritária durante a etapa de treinamento da base (MALDONADO; LÓPEZ; VAIRETTI, 2018).

Por fim, com a base já balanceada pode ser realizado o processo de normalização dos dados, a fim de possibilitar a aplicação de alguns algoritmos de aprendizado de máquina como, por exemplo, as redes neurais artificiais. Isso porque, os métodos de normalização permitem, por exemplo, converter os atributos para uma mesma escala. Esse processo de transformação de dados possuí algumas técnicas dentre as quais pode-se citar a normalização pelo escore-z e a pelo escalonamento decimal (DE CASTRO; FERRARI, 2016).

A técnica da normalização pelo escore-z é calculada a partir dos valores da média e do desvio padrão de um valor x. O resultado z obtido por esse método está representado na Equação 1, em que a média é representada por u e o desvio padrão por s (DE CASTRO; FERRARI, 2016).

$$z = (x - u) / s \quad (1)$$

Já o cálculo da técnica de normalização pelo escalonamento decimal, depende da quantidade de números decimais que serão movidos a partir do valor máximo de um atributo x. Para a obtenção do resultado é necessária a utilização de j que representa o menor valor inteiro desde que seja satisfeita a condição de $\max(|z|) < 1$. A Equação 2 representa o cálculo dessa normalização (DE CASTRO; FERRARI, 2016).

$$z = x/10^{j} \quad (2)$$

3.5 TRABALHOS CORRELATOS

Nesta seção são descritos os trabalhos relacionados ao presente estudo. Inicialmente, são apresentadas as pesquisas com a aplicação de algumas técnicas de aprendizado de máquina, com exceção da MLP, para a análise de dados com foco na vazão. Posteriormente, são apresentados estudos em que há a utilização, especificamente, da MLP para realizar a predição da mesma temática.

3.5.1 Utilização de aprendizado de máquina em vazão hídrica

Meshram et. al (2018) utilizou uma metodologia híbrida para tentar realizar a predição da vazão hídrica de uma bacia hidrográfica americana. Nessa pesquisa, foi utilizada a junção da rede neural *feed-foward* com o algoritmo híbrido *particle swarm*

optimization and gravitational search (PSOGSA). Para a criação desse artigo foram utilizados dados coletados entre 1990 e 2016 do rio Turkey do Estado de Iowa. Ao final do trabalho, os autores chegaram à conclusão de que, a partir dos resultados obtidos, a rede apresentou uma ótima eficiência na área estudada. Isso porque, o Erro médio quadrático da rede apresentou um valor de 24,42 m^3/s para o modelo estudado enquanto outros modelos como a FNN obteve um resultado de 33,58 m^3/s.

Já em Kalteh (2012), o autor fez a comparação entre a utilização de uma rede neural regular, o modelo *support vector regression* (SVR) e a adição em cada uma das duas de um transformador com *wavelet*s. Esse estudo foi realizado no Irã nas estações Kharjegil e Ponel com os dados do período de outubro de 1966 a setembro de 2006 e em uma base mensal relacionada à vazão. Ao final das análises, os pesquisadores chegaram à conclusão de que os modelos com *wavelets* possuem uma maior probabilidade de obterem um resultado com uma maior acurácia em relação aos que não empregaram esse transformador.

Em Le et al. (2019), os pesquisadores utilizaram a rede neural Long Short-Term Memory (LSTM) com o objetivo de tentar prever a possibilidade de alagamentos no Vietnã. Para esse estudo, os autores utilizaram dados do período de 1961 a 1984 para executarem o modelo. No caso desse artigo esses registros foram coletados do rio Da que pertence a uma das maiores bacias hidrográficas do país. Ao final da pesquisa os estudiosos chegaram à conclusão de que a LSTM trouxe resultados com uma alta precisão em determinados locais analisados. Porém, à longo prazo eles recomendaram que seja feita uma combinação com modelos meteorológicos para se alcançar resultados mais precisos para essa forma de avaliação.

3.5.2 Utilização da MLP em vazão hídrica

No estudo de Toro et. al (2013), os autores utilizaram um modelo híbrido de métodos da área da inteligência artificial e da estatística com o intuito de tentar realizar a predição do volume de água diário em uma bacia hidrográfica colombiana. Para a realização desse estudo foram utilizados dados do período de 1950 a 2006. Essas informações foram coletadas na bacia hidrográfica do rio Salvajina. Ao final do estudo, os pesquisadores chegaram à conclusão de que o método empregado obteve resultados expressivos com uma média absoluta do erro de 17,1141. A MLP, em comparação, obteve uma média absoluta do erro bem superior de 55,3210 ratificando

a eficiência do estudo. Com isso, chegaram à conclusão de que o modelo, que inicialmente foi utilizado para um estudo de curto prazo, pode ser utilizado também para pesquisas de longo prazo.

Por fim, em Ghorbani (2016), os autores realizaram um estudo comparativo entre a rede *multilayer perceptron* (MLP), uma *radial basis function* (RBF) e a *support vector machine* (SVM) a fim de predizer o valor da vazão do rio Zarrinehru. Os dados foram coletados no período de 1989 a 2011 no Irã. Ao final dos estudos, os pesquisadores chegaram à conclusão de que os três métodos podem ser utilizados na realização de pesquisas no campo da predição. Entretanto, as análises realizadas demonstraram que tanto a MLP quanto a RBF apresentaram melhores resultados quando comparadas ao modelo SVM. Isso porque, a MLP apresentou o critério de correlação médio e o erro quadrático médio de 0,813 e 11,947 m^3/s, respectivamente. Já a RBF apresentou os valores de 0,830 e 11,333 m^3/s. E, por fim, a SVM apresentou os valores de 0,790 e 13,840 m^3/s.

4. METODOLOGIA

Foram utilizadas as bases de dados CAPES e os sites da Agência Nacional de Águas e Saneamento Básico, Ministério da Agricultura e Google Acadêmico. Os artigos foram selecionados na língua portuguesa e inglesa, nos últimos dez anos, referentes à temática. Descritores utilizados: Água, *Data Science*, Mineração de dados, Dados ambientais, *Machine Learning*.

Para a realização desse estudo, os dados foram obtidos no site do Departamento de Águas e Energia Elétrica (DAEE). Na página, foi selecionado Banco de Dados “Fluviométricos” e a pesquisa foi realizada por Município. Além disso, no campo “Tipo de Dados” a opção selecionada foi a de “Vazões Médias Diárias”. Por fim, os dados foram filtrados pelo período desejado. Um recorte inicial da base está ilustrado na Figura 2.

Com o intuito de utilizar de informações mais recentes, os dados coletados pertencem ao intervalo de tempo de 2011 a 2020. Como, no site há uma grande quantidade de possibilidades de municípios e nem todos possuem registros do período desejado, para esse trabalho foi selecionada uma base da região de Campinas.

Figura 2 – Recorte dos dados da base de dados do Departamento de Águas e Energia

Mês/Ano	1	2	3	4	5	6	7	8	9	10
01/2010	67,226	101,709	116,819	105,026	92,228	115,458	130,283	121,271	110,719	99,073
02/2010	152,106	148,266	133,788	110,719	91,260	86,451	77,946	69,645	67,527	67,829
03/2010	145,213	105,026	82,962	70,556	60,676	59,209	57,750	52,566	53,708	47,773

Fonte: (DEPARTAMENTO DE ÁGUAS E ENERGIA ELÉTRICA, 2021)

Para a atual pesquisa as bases de dados utilizadas estão estruturadas da seguinte maneira: as vazões médias diárias, calculadas em m^3/s estão distribuídas por mês e ano que correspondem às linhas da tabela e dias do mês que se referem às colunas, conforme ilustrado na Figura 2. Além disso, é possível encontrar para todo mês/ano a informação da vazão média, mínima e máxima do período.

O modelo que será utilizado para realizar a predição é a MLP. Para a aplicação dessa rede, as bases foram rotuladas e divididas de forma com que 80% dos registros

foram utilizados durante o processo de treinamento enquanto os dados restantes foram usados para a fase de testes.

Em relação a rede, a MLP é comumente utilizada para estimação, entretanto, ela pode ser utilizada para resolver problemas de classificação. Com isso, no presente trabalho, apesar de a base ser constituída por números contínuos, os objetos foram rotulados com o intuito de obter os resultados divididos por classes e verificar os resultados obtidos pela MLP com a base separada por classes.

Para a criação do modelo, no caso da base mensal, foram utilizadas 28 entradas correspondentes aos registros de cada um dos 28 primeiros dias de cada mês. Já, no caso da diária, foi utilizada uma única entrada na rede correspondente ao valor do dia em que foi obtido o registro.

Ou seja, uma entrada do modelo mensal pode ser exemplificada por [43,906; 31,241; 31,493; ...; 40,113] em que cada um dos itens dessa lista, nesse caso, corresponde a um dia do mês de fevereiro de 2011. Já a saída esperada, nesse caso, é a classe “BAIXO”.

Já uma entrada do modelo diário pode ser representada por um valor, por exemplo, 32,258, que representa o dia 2 de janeiro de 2011. Já a saída esperada, nesse caso, é a classe “BAIXO”.

Com isso, foi criado um modelo que busca predizer a classe que será obtida no mês seguinte ao último registro da base mensal e um outro modelo que possuí o intuito de prever o rótulo do dia seguinte ao último da base diária.

Por fim, foi aplicada a técnica da validação cruzada para a obtenção da acurácia média após a execução do modelo 10 vezes.

4.1 TRATAMENTO DA BASE DE DADOS

As bases de dados coletadas, inicialmente, continham alguns dados faltantes, conforme Figura 3, que dificultariam a execução do modelo. Para evitar um possível enviesamento da base, as colunas referentes aos dias 29, 30 e 31 foram removidas visto que o valor desse atributo é irregular entre os objetos e que todas as colunas referentes a dias do mês foram utilizadas como entrada da rede. Por conta disso, a grande quantidade de registros ausentes nestes atributos poderia gerar uma

inconsistência na criação do modelo posteriormente. O resultado dessa limpeza está ilustrado na Figura 4.

Figura 3: Base de dados mensal com dados faltantes

23	24	25	26	27	28	29	30	31
73,308	63,040	60,088	49,734	45,830	40,650	37,453	38,778	34,574
33,798	47,773	58,624	41,729	35,354	40,113	---	---	---
26,511	26,756	26,023	24,814	23,617	24,573	26,267	39,311	37,717
23,142	23,142	22,905	22,905	23,142	28,979	29,478	29,979	---
17,844	16,955	17,844	15,858	15,858	17,176	16,734	16,076	16,076
17,882	17,882	17,468	18,299	18,299	17,675	18,508	18,508	---

Fonte: (DEPARTAMENTO DE ÁGUAS E ENERGIA ELÉTRICA, 2021)

Figura 4: Base de dados mensal após o processo de limpeza

20	21	22	23	24	25	26	27	28
112,406	104,693	95,148	73,308	63,04	60,088	49,734	45,83	40,65
48,332	43,359	38,778	33,798	47,773	58,624	41,729	35,354	40,113
35,094	32,258	26,756	26,511	26,756	26,023	24,814	23,617	24,573
25,538	25,055	23,379	23,142	23,142	22,905	22,905	23,142	28,979
19,196	18,292	18,067	17,844	16,955	17,844	15,858	15,858	17,176

Fonte: (DEPARTAMENTO DE ÁGUAS E ENERGIA ELÉTRICA, 2021)

A fim de realizar a técnica do encaixotamento de mesma largura apresentada por De Castro e Ferrari (2016) e com o intuito de categorizar a base, inicialmente, foi calculado o valor da média de cada objeto e, posteriormente, foram retirados o maior e o menor valor dos registros, sem a utilização dos dias 29, 30 e 31. A partir da diferença entre estes valores, realizou-se a criação de três conjuntos de caixas para que se encaixassem nas categorias “ALTA”, “MEDIO” e “BAIXO” com a seguinte distribuição: a primeira caixa com os valores entre 0,215 e 92,036, a segunda entre 92,037 e 183,859 e, por fim, a terceira entre 183,859 e 275,68.

Com base nesses resultados obtidos, foi feita a classificação dos resultados da seguinte forma: se o valor da média do registro estivesse contido no conjunto 3, foi atribuído a classe de “ALTA”. Caso, esse valor pertencesse à segunda caixa, a classe recebida foi “MEDIO”. Por fim, os dados pertencentes ao primeiro grupo foram classificados como “BAIXO”, conforme Figura 5.

Figura 5: Resultado da discretização da base mensal

16	17	18	19	20	21	22	23	24	25	26	27	28	
179,275	177,268	155,584	126,452	112,406	104,693	95,148	73,308	63,04	60,088	49,734	45,83	40,65	ALTO
37,453	85,496	69,645	59,795	48,332	43,359	38,778	33,798	47,773	58,624	41,729	35,354	40,113	BAIXO
28,73	26,756	26,023	36,4	35,094	32,258	26,756	26,511	26,756	26,023	24,814	23,617	24,573	MEDIO
38,247	32,258	29,728	27,001	25,538	25,055	23,379	23,142	23,142	22,905	22,905	23,142	28,979	BAIXO
20,339	20,569	20,339	19,651	19,196	18,292	18,067	17,844	16,955	17,844	15,858	15,858	17,176	BAIXO

Fonte: Arquivo Pessoal

Por conta do desbalanceamento da base, com a classe “BAIXO” apresentando uma quantidade de aproximadamente 93%, a “MEDIO” de aproximadamente 6% e a “BAIXO” de aproximadamente 1%, foi realizado um processo de *oversampling* na base com o intuito equiparar a quantidade de cada rótulo.

Além disso, com o intuito de criar, também, um modelo de predição à curto prazo, foi realizada a transposição da base transformando-a em diária. Nesse caso, a entrada da rede foi uma única coluna em que cada objeto correspondia a um dia de um ano. Ou seja, os atributos que consistiam nos meses do ano, foram reformulados para um único atributo com a informação do dia do registro, conforme Figura 6.

Figura 6: Resultado da transposição da base mensal para diária

09.01.2011	122.648
10.01.2011	87.727
11.01.2011	139.793
12.01.2011	147.883
13.01.2011	222.731

Fonte: Arquivo Pessoal.

Em seguida, assim como na base mensal, foi realizada a criação de 3 conjuntos de caixa para a realização da técnica de discretização, com o intuito de classificar os dados. Nesse caso, os intervalos ficaram da seguinte forma: a primeira caixa com os valores de 0,215 a 92,036, a segunda com os números entre 92,037 e 183,858 e, por fim, a terceira com o intervalo de 183,859 a 275,68.

Assim como anteriormente, foi realizada a classificação da base, dividindo as classes “BAIXO”, “MEDIO” e “ALTO” entre os conjuntos de caixa 1, 2 e 3, respectivamente nesta ordem. Entretanto, após uma análise, houve a necessidade de discretizar novamente a classe “BAIXO” pois esta apresentava a grande maioria dos registros. Assim sendo realizado o processo do encaixotamento outra vez.

Nesse caso, a distribuição das caixas ficou da seguinte maneira: a primeira caixa com os valores de 0,215 a 30,821, a segunda com os números entre 30,822 e 61,428 e, finalmente, a terceira caixa com o intervalo de 61,429 e 92,037. A partir dessa divisão, a classe “BAIXISSIMO” foi atribuída para a caixa um, o segundo conjunto foi rotulado como “BAIXO” e, por fim, à última foi dada a classe “BAIXO ALTO”, conforme Figura 7.

Por se tratarem de duas bases de dados distintas que foram utilizadas para a criação de dois modelos, a mesma metodologia foi aplicada na base de dados diária com a diferença de que, neste caso, o processo de encaixotamento foi realizado duas vezes. Portanto, foram gerados dois classificadores.

Figura 7: Resultado da discretização da base diária

09.01.2011	122,648	MEDIO
10.01.2011	87,727	BAIXO ALTO
11.01.2011	139,793	MEDIO
12.01.2011	147,883	MEDIO
13.01.2011	222,731	ALTO

Fonte: Arquivo Pessoal.

Um outro ponto é que, assim como na base mensal, as classes que representavam os valores menores, mesmo após a segunda discretização,

continuaram desbalanceados em relação às maiores. Com isso, foi utilizada a técnica SMOTE do *oversampling* novamente com o intuito de tornar a base mais equilibrada.

Por fim, para possibilitar a leitura dos arquivos na linguagem Python e no modelo criado, foram retirados os nomes dos atributos da tabela deixando, assim, apenas os registros. Além disso, os arquivos foram convertidos para o formato ".csv".

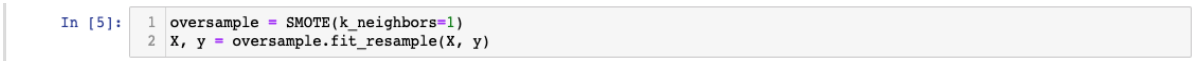
4.2 LINGUAGEM DE PROGRAMAÇÃO

Para a execução da MLP foi escolhida a linguagem de programação Python e, para isso, foram utilizadas as bibliotecas "pandas", "sklearn" e "imblearn". O código utilizado foi baseado em Leonel (2019). Nesse estudo realizado pelo autor, a rede neural foi criada com o intuito de classificar diferentes tipos de plantas do gênero Íris.

No código, primeiramente, foi feita a leitura dos arquivos .csv que continham as bases a serem estudadas. Em seguida, foi criado um vetor com o nome atrelado a cada coluna da tabela. Posteriormente, foi atribuída a variável X os registros das 28 primeiras colunas, no caso da base mensal, e a primeira coluna, no caso da diária e à Y a classe da respectiva tupla. As classes, em seguida, foram transformadas em valores numéricos que iam do valor 0 a 2, no primeiro caso, e de 0 a 4, no caso da análise a curto prazo.

Em ambos os casos, foi realizado o processo de oversampling das bases de dados a partir da técnica SMOTE, que consiste basicamente em gerar novas objetos das classes minoritárias a partir da interpolação de instâncias escolhidas a partir de k vizinhos mais próximos (MALDONADO; LÓPEZ; VAIRETTI, 2018). No caso do modelo mensal, para o funcionamento correto do método, foi definido o número de k igual a 1, conforme Figura 8. Isso porque, a classe minoritária da base, no caso "ALTA", apresenta apenas um registro associado a ela. Já no caso da diária, foi utilizado o valor de k como 5 já que este número corresponde ao valor padrão do método na biblioteca utilizada.

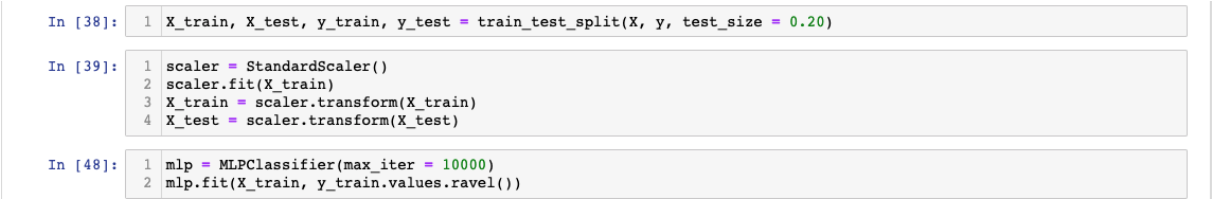
Figura 8: Aplicação da técnica de *oversampling*



Fonte: Arquivo Pessoal.

Após essas atribuições, deu-se início ao processo de treinamento e de testes dos modelos e, para esse estudo, foi utilizado 80% da base para a etapa de treinamento e o restante para os testes. Em seguida, foi realizada a normalização da base a partir da técnica escore-z com a utilização do método StandardScaler. Por fim, com o intuito de possibilitar que as redes sempre convergissem, foi determinado o valor máximo de iterações como 10000 e foi executado o modelo com o MLPClassifier da linguagem, conforme a Figura 9.

Figura 9: Treinamento e criação do modelo



Fonte: Arquivo Pessoal.

Com os modelos criados, foi realizado um estudo para a definição da quantidade de neurônios na camada intermediária das redes, sendo utilizado o estimador GridSearchCV e, em ambos os casos, foi passada a métrica acurácia e um objeto com a chave sendo a quantidade de neurônios, que é o objetivo a ser descoberto, e uma lista com as quantidades de 1 a 20 para a determinação do melhor resultado para os modelos. Após isso, foram selecionadas as quantidades que possuíam melhor acurácia média. Para esse estudo da determinação da quantidade de neurônios, os resultados foram obtidos a partir do subconjunto de treinamento.

Por fim, foram criados os modelos com as quantidades de neurônios na camada intermediária encontradas pelo estimador. Com as MLP's criadas, foram realizadas as validações cruzadas para a obtenção das médias das acurácias com o intuito de analisar a eficiência dos modelos em relação aos problemas estudados.

5. RESULTADOS

Em relação à base mensal, na execução do SMOTE, inicialmente obteve-se o aumento do número de registros de 125 para 345. Na sequência, foi treinado o estimador e o resultado obtido foi de que a melhor acurácia média para o modelo proposto foi obtida com a utilização de 9 neurônios na camada intermediária, conforme ilustrado na Tabela 1.

Tabela 1: Relação entre número de neurônios na camada intermediária e acurácia média para a base mensal

Número de neurônios	Acurácia Média
1	0,455072
2	0,521739
3	0,811594
4	0,582609
5	0,939130
6	0,907246
7	0,921739
8	0,944928
9	0,973913
10	0,944928
11	0,846377
12	0,973913
13	0,617391
14	0,953623
15	0,971014
16	0,971014
17	0,947826
18	0,971014
19	0,968116
20	0,971014

Fonte: Arquivo Pessoal.

Após isso, foi realizada a execução do modelo com o melhor número de neurônios na camada intermediária que foi o de 9 neurônios. Com base nesse valor, foi realizada a validação cruzada com o número de k sendo igual a 10. O resultado obtido foi que a acurácia média para essa MLP foi de aproximadamente 95%.

Para a base diária, a execução do SMOTE resultou no aumento na quantidade de registros de 3500 a 14725. Após isso, o estimador foi novamente utilizado para

calcular a melhor quantidade de neurônios na camada intermediária. O resultado obtido foi que, neste caso, o número ideal seria a utilização de 11 neurônios, conforme ilustrado na Tabela 2.

Tabela 2: Relação entre número de neurônios na camada intermediária e acurácia média para a base diária

Número de neurônios	Acurácia Média
1	0,324075
2	0,800679
3	0,954975
4	0,988183
5	0,987436
6	0,990153
7	0,988998
8	0,989406
9	0,990628
10	0,986078
11	0,992190
12	0,991647
13	0,989202
14	0,983837
15	0,990560
16	0,990221
17	0,990900
18	0,988183
19	0,991511
20	0,982139

Fonte: Arquivo Pessoal.

Posteriormente, foi realizada novamente a execução do modelo com a quantidade de neurônios na camada intermediária sendo o melhor número encontrado anteriormente, no caso, o valor de 11. Outra vez foi feita a validação cruzada com o número de k sendo igual a 10. O resultado obtido foi que a acurácia média para essa MLP foi de aproximadamente 99%.

Para realizar uma comparação, foi executado o modelo novamente com as mesmas especificações, tanto para o caso da base diária quanto para a base mensal, porém, desta vez, sem a utilização do método de oversampling. Em relação à base mensal, a acurácia média após o processo de validação cruzada foi de aproximadamente 80%, enquanto para a diária o valor foi, assim como anteriormente, de aproximadamente 99%.

6. CONCLUSÃO

Neste trabalho foi utilizada a rede neural MLP com o intuito de criar um modelo preditivo para o problema da vazão diária. Para isso, foi utilizada uma base de dados da região de Campinas que, inicialmente, possuí os dados representados de forma mensal e, posteriormente, foram alteradas também para a criação de um modelo diário.

Essa divisão foi realizada com o intuito de criar um modelo que contemplasse t um planejamento à curto prazo, no caso do modelo diário, e um à médio prazo, no caso do mensal. Com isso, há um auxílio tanto no processo de estudo a menor prazo, com o intuito de prevenir possíveis problemas imediatos, como enchentes, quanto na ajuda em um melhor planejamento em relação a um maior intervalo de tempo.

Em relação a base de dados mensal, o modelo criado apresentou uma melhor acurácia média após o balanceamento das classes, a partir da técnica de *oversampling*, apresentando uma melhora de 15% quando comparada a base desbalanceada.

Já no caso da base de dados diária, não houve diferença com a utilização da técnica de *oversampling*, com ambos os casos apresentando uma boa acurácia média para o problema proposto para ser estudado. Isso se deve ao fato da base apresentar um número já considerável de registros antes mesmo da técnica de *oversampling.*

A partir desse trabalho, é possível perceber que estudos em relação a esse problema com a utilização de métodos preditivos ainda são bem escassos no território brasileiro. Uma melhora nesse aspecto, pode auxiliar em um melhor planejamento e, consequentemente, uma diminuição em problemas relacionados à vazão de rios.

REFERÊNCIAS BIBLIOGRÁFICAS

AGÊNCIA NACIONAL DE ÁGUAS E SANEAMENTO BÁSICO. **Quantidade de água**. Brasil. Disponível em: https://www.ana.gov.br/aguas-no-brasil/panorama-das-aguas/quantidade-da-agua. Acesso em: 24 ago. 2020.

AKINER, Muhammed Ernur; AKKOYUNLU, Atila. Modeling and forecasting river flow rate from the Melen Watershed, Turkey. **Journal of Hydrology**. v. 456-457. p. 121-129. 2012. Disponível em: https://www.sciencedirect.com/science/article/abs/pii/S0022169412005161?via%3Dihub. Acesso em: 25 jun. 2021

ALI, Maqbool et al. A Data-Driven Knowledge Acquisition System: An End-to-End Knowledge Engineering Process for Generating Production Rules. **IEEE Access**. v. 6, p. 15587-15607, mar. 2018. Disponível em: https://ieeexplore.ieee.org/document/8319403. Acesso em: 24 ago. 2020.

AMBROSIO, Julia Kobylanski. **Comitê de máquinas para previsão da demanda de água**. 2019. Tese (Mestrado em Tecnologia na especialidade Ambiente) – Universidade Estadual de Campinas, Limeira, 2019.

BRENTAN, B.; MEIRELLES, G. et al. Correlation analysis of water demand and predictive variables for short-term forecasting models. Mathematical Problems in Engineering, v.2017, p. 1-10, 2017. Disponível em: https://downloads.hindawi.com/journals/mpe/2017/6343625.pdf. Acesso em: 25 mar. 2021.

DE MARCHIS, Mauro et al. Energy Saving in Water Distribuition Network through Pump as Turbine Generators: Economic and Environmental Analysis. **Energies.** v. 9, p. 877, Out. 2016. Disponível em: https://www.researchgate.net/publication/309472033_Energy_Saving_in_Water_Distribution_Network_through_Pump_as_Turbine_Generators_Economic_and_Environmental_Analysis. Acesso em: 16 dez. 2020.

BOU-FAKHREDDINE, Bassam et al., Daily river flow prediction based on Two-Phase Constructive Fuzzy Systems Modeling: A case of hydrological – meteorological measurements asymmetry. **Journal of Hydrology**. v. 558. p. 255-265. Mar. 2018. Disponível em: https://www.sciencedirect.com/science/article/abs/pii/S0022169418300350?via%3Dihub. Acesso em: 25 jun. 2021.

DE CASTRO, Leandro Nunes; FERRARI, Daniel Gomes. **Introdução à Mineração de Dados**. 1. ed. São Paulo: Saraiva, 2016.

DEPARTAMENTO DE ÁGUAS E ENERGIA ELÉTRICA. **Banco de Dados Hidrológicos**. Brasil. Disponível em: http://www.hidrologia.daee.sp.gov.br/. Acesso em: 25 jun. 2021.

GEDEFAW, Mohammed et al., Variable selection methods for water demand forecasting in Ethiopia: Case study Gondar town. **Cogent Environmental Science**, v. 4, p. 1-11, Out. 2018. Disponível em: https://www.tandfonline.com/doi/pdf/10.1080/23311843.2018.1537067?needAccess=true. Acesso em: 13 maio 2021.

GHORBANI, Mohammad Ali et al. A comparative study of artificial neural network (MLP, RBF) and support vector machine models for river flow prediction. **Environmental Earth Sciences**. v. 75. n. 476. Mar. 2016. Disponível em: https://link.springer.com/article/10.1007%2Fs12665-015-5096-x. Acesso em: 25 jun. 2021.

GHORBANI, Saeed; BARARI, Morteza; HOSEINI, MOJTABA. Presentig a new method to improve the detection of micro-seismic events. **Environmental Monitoring and Assessment**. V.190, p. 464, Jul. 2018. Disponível em: https://link.springer.com/article/10.1007/s10661-018-6837-6. Acesso em: 17 nov. 2021.

GILBERT, Karina; SÀNCHEZ-MARRÈ, Miquel; IZQUIERDO, Joaquin. A survey on pre-processing techniques: Relevant issues in the context of environmental data mining. **AI communications**, v. 29, n. 6, p. 627-663, Dez. 2016. Disponível em: https://www.cs.upc.edu/~idss/transpasK/AIC710def.pdf. Acesso em: 24 ago. 2020.

HAQUE, Md Mahmudul et al., A comparative assessment of variable selection methods in urban water demand forecasting. **Water**. v. 10, n. 4, p. 419, Abr. 2018. Disponível em: https://www.mdpi.com/2073-4441/10/4/419. Acesso em: 13 maio 2021.

KALTEH, Aman Mohammed. Monthly river flow forecasting using artificial neural network and support vector regression models coupled with wavelet transform. **Computers & Geosciences**. v. 54. p. 1-8. Abr. 2013. Disponível em: https://www.sciencedirect.com/science/article/pii/S0098300412003925. Acesso em: 25 jun. 2021.

KAMIŃSKI, Kamil; KAMIŃSKI, Wladyslay; MIZERSKI Tomasz. Application of artificial neural networks to the technical condition assessment of water supply systems. **Ecological Chemistry and Engineering S**. v. 24, n. 1, p. 31-40, Fev. 2017. Disponível em: https://www.researchgate.net/publication/316068171_Application_of_Artificial_Neural_Networks_to_the_Technical_Condition_Assessment_of_Water_Supply_Systems. Acesso em: 16 dez. 2020.

LE, Xuan-Hien et al., Application of Long-Short-Term Memory (LSTM) Neural Network for Flood Forecasting. **Water**. v. 11. n. 7. p. 1387, Jul. 2019. Disponível em: https://www.mdpi.com/2073-4441/11/7/1387. Acesso em: 25 jun. 2021.

LEONEL, Jorge. **MultiLayer-Perceptron**. Brasil, 2019. Disponível em: https://github.com/jorgesleonel/Multilayer-Perceptron. Acesso em: 17 nov. 2021.

LORENTE-LEYVA, L. Leandro et al., Artificial Neural Networks for Urban Water Demand Forecasting: A Case Study. **Journal of Physics: Conference Series**. p. 1-

8, Jul. 2019. Disponível em: https://www.researchgate.net/publication/335322004_Artificial_Neural_Networks_for_Urban_Water_Demand_Forecasting_A_Case_Study. Acesso em: 25 mar. 2021.

MALDONADO, Sebastián; LÓPEZ, Julio; VAIRETTI, Carla. An Alternative SMOTE Oversampling Strategy for High-dimensional Datasets. **Applied Soft Computing Journal.** v.76. p. 380-389, Dec. 2018. Disponível em: https://www.sciencedirect.com/science/article/abs/pii/S1568494618307130. Acesso em: 7 nov. 2021.

MANZAN, Jose Ricardo Gonçalves. Análise de desempenho de redes neurais artificiais do tipo Multilayer Perceptron por meio de distanciamento dos pontos do espaço de saída. 2016. Tese (Doutorado em Ciências) – Universidade Federal de Uberlândia, Uberlândia, 2016.

MEDEIROS, R. et al., **Previsão de demanda a médio prazo aplicada em dados reais do sistema de distribuição: uma comparação entre RNA e Lógica Fuzzy.** Divulgação cientifica e tecnológica do IFPB., 2016. Disponível em: https://www.researchgate.net/publication/311854765_Previsao_de_demanda_a_medio_prazo_aplicada_em_dados_reais_do_sistema_de_distribuicao_uma_comparacao_entre_RNA_e_Logica_Fuzzy. Acesso em: 16 dez. 2020.

MESHRAM, Sarita Gajbhiye et al. River flow prediction using hybrid PSOGSA algorithm based on feed-forward neural network. **Soft computing**. v. 23. p. 10429-10438. Nov. 2019. Disponível em: https://link.springer.com/article/10.1007/s00500-018-3598-7. Acesso em: 25 jun. 2021.

MINEMATO, Toshifume et al. Feed forward neural network with random quaternionic neurons. **Signal Processing**. Japão, v. 136, p. 59-68, Nov. 2016. Disponível em: https://www.researchgate.net/publication/310390375_Feed_forward_neural_network_with_random_quaternionic_neurons. Acesso em: 09 abr. 2021

POLI, Claudia Maria Basso. As causas e formas de prevenção sustentáveis das enchentes urbanas. In: Seminário Nacional de Construções Sustentáveis, 2, 2013, Passo Fundo, Brasil. IMED, 2013. p. 1-7.

RAHIN, Md Shamsur et al. Machine Learning and Data Analytic Techniques in Digital Water Metering: A Review. **WATER**, v. 12, n. 1, p.1-26, Jan. 2020. Disponível em: https://www.mdpi.com/2073-4441/12/1/294/htm. Acesso em: 24 ago. 2020.

SATHYA, R; ABRAHAM, Annamma. Comparison of Supervised and Unsupervised Learning Algorithms for Pattern Classification. **International Journal of Advanced Research in Artificial Intelligence**. v.2, p. 34-38, Fev. 2013. Disponível em: https://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.278.5274&rep=rep1&type=pdf#page=41. Acesso em: 9 abr. 2021.

SEBRI, Maamar. Forecasting urban demand: A meta-regression analysis. **Journal of Environmental Management**. Tunisia. v. 183, p. 777-785, Dez. 2016 Disponível em: https://www.sciencedirect.com/science/article/pii/S0301479716306909. Acesso em: 25 mar. 2021.

SEBRI, Maamar. ANN versus SARIMA models in forecasting residential water consumption in Tunisia. **Journal of Water, Sanitation and Hygiene for Development**. v. 3, n. 3, p. 330-340, Fev. 2013. Disponível em: https://iwaponline.com/washdev/article/3/3/330/30305/ANN-versus-SARIMA-models-in-forecasting. Acesso em: 25 mar. 2021.

SEO, Youngmin; KWON, Soonmyeong; CHOI, Yunyoung. Short-Term Water Demand Forecasting Model Combining Variational Mode Decomposition and Extreme Learning Machine. **Hydrology**. Basel. v. 5, n. 4, p. 44, Set. 2018. Disponível em: https://www.researchgate.net/publication/327933620_Short-Term_Water_Demand_Forecasting_Model_Combining_Variational_Mode_Decomposition_and_Extreme_Learning_Machine. Acesso em: 25 mar. 2021

SHARMA, Vidushi; RAI, Sachin; DEV, Anurag. A Comprehensive Study of Artificial Networks. **International Journal of Advanced Research in Computer Science and Software Engineering**. v. 2, p. 278-284, Out. 2012. Disponível em: https://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.468.9353&rep=rep1&type=pdf. Acesso em: 16 dez. 2020.

SOUZA, Nayara Silva; SOUZA, Wanderley de Jesus; CARDOSO, Jossy Mara Simões. Caracterização hidrológica e influência da cobertura do solo nos parâmetros de vazão do Rio das Fêmeas. **Engenharia Sanitária e Ambiental**. v. 22. n. 3. p. 453-462. Jun. 2017. Disponível em: http://www.scielo.br/j/esa/a/jKBdnLybP3rCWK7rdfmftPy/abstract/?lang=pt. Acesso em: 25 jun. 2021.

SRIVASTAVA, Shweta. Weka: A Tool for Data preprocessing, Classification, Ensemble, Clustering and Association Rule Mining. **International Journal of Computer Applications**, v. 88, n. 10, p. 26-29, Fev. 2014. Disponível em: http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.429.1463&rep=rep1&type=pdf. Acesso em: 24 ago. 2020.

SUH, Dongjun; HAM, Sungil. A water demand forecasting model using BPNN for residential building. **Contemporary Engineering Sciences**. v. 9, n. 1, p. 1-10, Dec. 2015. Disponível em: https://www.researchgate.net/publication/298713886_A_water_demand_forecasting_model_using_BPNN_for_residential_building. Acesso em: 09 abr. 2021

THAMILSELVAN, P; SATHIASEELAN, Dr. J. G. R. **International Journal of Education and Management Engineering**. v.2, p. 1-9, Jun. 2015. Disponível em: https://www.researchgate.net/publication/281666263_A_Comparative_Study_of_Data_Mining_Algorithms_for_Image_Classification. Acesso em: 16 dez. 2020.

TIWARI, Mukesh; ADAMOWSKI, Jan; ADAMOWSKI, Kazimierz. Water demand forecasting using extreme learning machines. **Journal of Water and Land Development**, Polônia, v. 28, n. 1, p. 37-52, Maio 2016. Disponível em: https://content.sciendo.com/view/journals/jwld/28/1/article-p37.xml. Acesso em: 21 ago. 2020.

TIWARI, K. Mukesh; ADAMOWSKI, Jan. Urban water demand forecasting and uncertainty assessment using ensemble wavelet-bootstrap-neural network models. **Water Resources Research,** Polônia, v. 48, p. 6486-6507, Out. 2013. Disponível em: https://agupubs.onlinelibrary.wiley.com/doi/epdf/10.1002/wrcr.20517. Acesso em: 25 mar. 2021.

TORO, Carlos H. Fajardo et al. A hybrid artificial intelligence model for river flow forecasting. **Applied Soft Computing**. v. 13, n. 8, p. 3449-3458, Aug. 2013. Disponível em: https://www.sciencedirect.com/science/article/abs/pii/S1568494613001385?via%3Dihub. Acesso em: 25 jun. 2021.

WANG, Yuan; WANG, Yao; LUI, Yvonne W. Generalized Recurrent Neural Network accommodating Dynamic Causal Modelling for functional MRI analysis. **NeuroImage**. v.178, p. 385-402, Maio 2018. Disponível em: https://www.researchgate.net/publication/325241135_Generalized_Recurrent_Neural_Network_accommodating_Dynamic_Causal_Modeling_for_functional_MRI_analysis. Acesso em: 09 mar. 2021.

XU, Yuebing et al. A Novel Dual-Scale Deep Belief Network Method for Daily Urban Water Demand Forecasting. **Energies**. v. 11, n. 5, p. 1065, Abr. 2018. Disponível em: mdpi.com/1996-1073/11/5/1068. Acesso em: 09 mar. 2021.

ZENG, Xublin et al. A toy model for monthly river flow forecasting. **Journal of Hydrology**. v. 452-453, p. 226-231, Jul. 2012. Disponível em: https://www.sciencedirect.com/science/article/abs/pii/S0022169412004568?via%3Dihub. Acesso em: 25 jul. 2021.

YASEEN, Zaher Mundher et al. Hourly River Flow Forecasting: Application of Emotional Neural Networks Versus Multiple Machine Learning Paradigms. **Water Resources Management**. v. 34, p. 1075-1091, Jan. 2020. Disponível em: https://link.springer.com/article/10.1007/s11269-020-02484-w. Acesso em: 25 jun. 2021.

ZHANG, Yuanyuan et al. Adaptive Convolutional Neural Network and It's Application in Face Recognition. **Neural Processing Letters**. Estados Unidos, v.43, p. 389-399, Mar. 2015. Disponível em: https://link.springer.com/article/10.1007/s11063-015-9420-y. Acesso em: 09 mar. 2021.