UNIVERSIDADE DE SÃO PAULO
Instituto de Ciências Matemáticas e de Computação

Modelo de classificação multivariável para identificação de enchentes: um estudo empírico no sistema de monitoramento de rios e-noe

Lucas Augusto Vieira Brito

Dissertação de Mestrado do Programa de Pós-Graduação em Ciências de Computação e Matemática Computacional (PPG-CCMC)

ICMC USP
SÃO CARLOS

Lucas Augusto Vieira Brito

Modelo de classificação multivariável para identificação de enchentes: um estudo empírico no sistema de monitoramento de rios e-noe

Dissertação apresentada ao Instituto de Ciências Matemáticas e de Computação – ICMC-USP, como parte dos requisitos para obtenção do título de Mestre em Ciências – Ciências de Computação e Matemática Computacional. *VERSÃO REVISADA*

Área de Concentração: Ciências de Computação e Matemática Computacional

Orientador: Prof. Dr. Jó Ueyama

USP – São Carlos
Julho de 2019

Ficha catalográfica elaborada pela Biblioteca Prof. Achille Bassi
e Seção Técnica de Informática, ICMC/USP,
com os dados inseridos pelo(a) autor(a)

Brito, Lucas Augusto Vieira
B862m Modelo de classificação multivariável para identificação de enchentes: um estudo empírico no sistema de monitoramento de rios e-noe / Lucas Augusto Vieira Brito; orientador Jó Ueyama. -- São Carlos, 2019.
70 p.

Dissertação (Mestrado - Programa de Pós-Graduação em Ciências de Computação e Matemática Computacional) -- Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo, 2019.

1. CRISP-DM. 2. RSSF. 3. Aprendizado de máquina. 4. Mineração de dados. 5. Identificação de enchentes. I. Ueyama, Jó , orient. II. Título.

Bibliotecários responsáveis pela estrutura de catalogação da publicação de acordo com a AACR2:
Gláucia Maria Saia Cristianini - CRB - 8/4938
Juliana de Souza Moraes - CRB - 8/6176

Lucas Augusto Vieira Brito

Multivariate classification model for identification of floods: an empirical study in the monitoring of e-noe rivers

Dissertation submitted to the Institute of Mathematics and Computer Sciences – ICMC-USP – in accordance with the requirements of the Computer and Mathematical Sciences Graduate Program, for the degree of Master in Science. *FINAL VERSION*

Concentration Area: Computer Science and Computational Mathematics

Advisor: Prof. Dr. Jó Ueyama

USP – São Carlos
July 2019

Este trabalho é dedicado a toda minha família, Brito e Deodato, por todo apoio e toda força que me deram nesses dois anos de estudos no mestrado e nos cinco anos de estudos na graduação, por me ajudarem a enfrentar todos os desafios e assim realizar o sonho de ser um pesquisador do Instituto de Ciências Matemáticas e de Computação (ICMC).

AGRADECIMENTOS

Primeiramente, gostaria de agradecer a Deus por me dar saúde e a oportunidade de estudar em uma das melhores universidades do Brasil, a USP (ICMC), e por me ajudar a percorrer todo caminho da Pós-Graduação com muita determinação e foco.

À minha família Brito, por me dar todo apoio aos estudos e ajudar a superar os desafios que a vida me proporcionou. Muito obrigado por tudo.

À minha família Deodato, agradeço pela confiança e pelo amor que me deram durante os anos que passei de graduação, me proporcionando todo o suporte para que eu fizessem uma ótima faculdade.

Aos meus amigos e companheiros do Intermídia, Heitor, Flávia, Neto, Geraldo, Felipe, Alef, Kishi, Mano, Sandra, Humberto, Thiago Costa, Tiago Trojahn, entre outros. Foi um prazer pesquisar ao lado de todos, muito obrigado por tudo.

À minha namorada e futura esposa Maria Izabel, por todo o apoio emocional e ajuda nas horas mais difíceis. Seu amor e companheirismo foram o combustível essencial para que eu pudesse alcançar todos os meus sonhos.

À UNIARA (Universidade de Araraquara), onde cursei a graduação e pude contar com uma ótima estrutura e excelentes professores, o que me impulsionou a realizar um dos meus maiores sonhos.

Ao Prof. Jó Ueyama, pelos ensinamentos tanto na orientação acadêmica quanto na vida e por toda ajuda em meu trabalho.

À Capes pelo suporte financeiro durante meu mestrado pelo processo PROEX-9916515/M.

“Suba o primeiro degrau com fé.
Não é necessário que você veja toda a escada. Apenas dê o primeiro passo.”
(Martin Luther King)

RESUMO

BRITO, V. L. A. **Modelo de classificação multivariável para identificação de enchentes: um estudo empírico no sistema de monitoramento de rios e-noe**. 2019. 70 p. Dissertação (Mestrado em Ciências – Ciências de Computação e Matemática Computacional) – Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo, São Carlos – SP, 2019.

Nas últimas décadas, as enchentes vêm causando muitos problemas nas cidades, principalmente em grandes centros urbanos devido à alteração da paisagem natural e à impermeabilização do terreno. Geralmente esses eventos estão relacionados a eventos extremos de chuva, junto a um insuficiente sistema de drenagem para dar vazão ao escoamento gerado. Um ponto agravante - que colabora com o aumento da magnitude das enchentes - é o crescimento populacional desordenado. Assim, faltam políticas públicas, como um estudo prévio da região para alocação de pessoas de maneira eficiente. Na literatura, existem algumas soluções, como o uso da tecnologia de Redes de Sensores Sem Fio (RSSF), que podem ser implantadas no cenário urbano como forma de monitoramento de enchentes. Nesse cenário, um dos principais desafios para elaboração desses sistemas é emitir alertas para que desastres maiores sejam evitados. Porém, a utilização de uma única fonte de dados, unida a possíveis falhas que as RSSFs podem sofrer, acaba comprometendo o monitoramento e o alerta de enchentes. Uma outra abordagem é a utilização de modelos hidrológicos criados a partir de um estudos prévios do solo e da estrutura da bacia, pois eles são capazes de reproduzir o comportamento do escoamento da bacia a partir de séries temporais como entrada. Existem muitos modelos hidrológicos com diversas estruturas de dados e detalhamento da bacia hidrográfica, dos mais complexos - capazes de reproduzir a física dos processos de infiltração e o escoamento de água - até os mais simplificados, que utilizam parâmetros de ajustes que não são necessariamente relacionados aos fenômenos físicos envolvidos nesses processos. Porém, muitos desses modelos precisam de uma grande quantidade de dados para o seu desenvolvimento, tornando-os muito complexos e custosos. Dessa forma, esta dissertação de mestrado apresenta um modelo de identificação de enchentes baseado na mineração de dados e aprendizado de máquina, com o intuito de diminuir a complexidade e o custo dos modelos hidrológicos e a dependabilidade de uma única variável de sistemas de RSSF, além da vantagem de ser facilmente generalizável sem perder a eficiência na identificação de enchente. As variáveis utilizadas para o desenvolvimento do modelo são os dados de estações meteorológicas e o nível de água do canal. Assim, é utilizada a metodologia do *Cross Industry Standard Process for Data Mining* (CRISP-DM) para a mineração dos dados, por ser uma técnica objetiva que contém as melhores práticas para a exploração dos dados. Os resultados revelam que o modelo desenvolvido obteve uma acurácia de aproximadamente 87.8%, com o algoritmo *Random_Forest*. Além disso, nos testes de adaptabilidade e comparação com o *Storm Water Management Model* (SWMM)-um modelo hidrológico amplamente conhecido na literatura-, em uma mesma região de estudo, o modelo desenvolvido obteve resultados relevantes no contexto de identificação de enchente. Isso

mostra que o modelo desenvolvido possui grande potencial de aplicação, principalmente por sua simplicidade de implementação e replicação sem comprometer a qualidade de identificação da ocorrência de enchentes. Consequentemente, algumas das principais contribuições deste trabalho são: (*i*) o modelo multivariável de identificação de enchente diminui a complexidade, custos e tempo de desenvolvimento em relação aos modelos hidrológicos e; (*ii*) o avanço do estado da arte em comparação aos trabalhos computacionais, por não depender de variáveis fixas e utilizar multivariáveis para identificar o padrão de enchentes.

Palavras-chave: CRISP-DM, RSSF, Aprendizado de máquina, Mineração de dados, Identificação de enchentes.

ABSTRACT

BRITO, V. L. A. **Multivariate classification model for identification of floods: an empirical study in the monitoring of e-noe rivers**. 2019. 70 p. Dissertação (Mestrado em Ciências – Ciências de Computação e Matemática Computacional) – Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo, São Carlos – SP, 2019.

In recent decades, floods have caused many problems in cities, especially in large urban centers due to the alteration of the natural landscape and the waterproofing of the terrain. Generally, these events are related to extreme rainfall events, together with an insufficient drainage system to give flow to the flow generated. An aggravating point - which contributes to the increase in flood magnitude - is disordered population growth. Thus, public policies are lacking, such as a prior study of the region for the efficient allocation of people. In the literature, there are some solutions, such as the use of the Wireless Sensor Networks (WSN) technology, which can be implemented in the urban scene as a form of flood monitoring. In this scenario, one of the major challenges in designing these systems is to issue alerts so that major disasters are avoided. However, the use of a single data source, coupled with the possible flaws that WSNs may suffer, endangers flood monitoring and alertness. Another approach is the use of hydrological models created from previous soil studies and basin structure, as they are able to reproduce basin flow behavior from time series as input. There are many hydrological models with diverse data structures and details of the hydrographic basin, of the most complex - capable of reproducing the physics of the infiltration processes and the water flow - to the more simplified, that use parameters of adjustments that are not necessarily related to the phenomena involved in these processes. However, many of these models need a lot of data for their development, making them very complex and costly. This dissertation presents a flood identification model based on data mining and machine learning in order to reduce the complexity and cost of hydrological models and the dependability of a single variable of WSN systems. of the advantage of being easily generalizable without losing efficiency in the identification of flood. The variables used for the development of the model are the data of meteorological stations and the water level of the channel. Thus, the Cross Industry Standard Process for Data Mining (CRISP-DM) methodology for data mining is used, since it is an objective technique that contains the best practices for data mining. The results show that the developed model obtained an accuracy of approximately 87.8%, with the algorithm Random_Forest. In addition, in the adaptive and comparative tests with the Storm Water Management Model (SWMM), a hydrological model widely known in the literature, in the same region of study, the developed model obtained relevant results in the context of flood identification. This shows that the developed model has great application potential, mainly for its simplicity of implementation and replication without compromising the quality of the identification of the occurrence of floods. Consequently, some of the main contributions of this work are: (i) the multivariate model of flood identification decreases the

complexity, costs and development time in relation to the hydrological models; (ii) the advance of the state of the art in comparison to the computational works, because it does not depend on fixed variables and use multivariable to identify the flood pattern.

Keywords: CRISP-DM, WSN, Machine learning, Data mining, flood identification.

LISTA DE ILUSTRAÇÕES

LISTA DE TABELAS

LISTA DE ABREVIATURAS E SIGLAS

MANET	*Mobile Ad hoc Network*
CRISP-DM	*Cross Industry Standard Process for Data Mining*
EESC	Escola de Engenharia de São Carlos
ICMC	Instituto de Ciência Matemática e de Computação
KNN	*K-Nearest Neighboor*
MLP	*Multilayer Perceptron*
RSSF	Redes de Sensores Sem Fio
SVM	*Support Vector Machines*
SWMM	*Storm Water Management Model*

SUMÁRIO

CAPÍTULO

1

INTRODUÇÃO

Os processos sociais, econômicos e culturais em desenvolvimento têm provocado um crescimento populacional acelerado e desorganizado em áreas urbanas (SILVA; PORTO, 2003). Segundo a (ONU, 2014), a população urbana representará dois terços da população mundial até 2050. No mesmo período, estima-se que a população rural diminuirá de 3,4 bilhões para 2,9 bilhões de pessoas.

O processo acelerado de urbanização que vem ocorrendo nas últimas décadas, atrelado a mudanças climáticas e ambientais globais, pode produzir impactos sobre a saúde humana por diferentes vias e intensidades. Os grandes centros urbanos serão as regiões mais afetadas pelos os inúmeros impactos ambientais que afetam a vida de milhões de pessoas no mundo. As enchentes aparecem de forma cada vez mais frequentes e intensas nessas regiões pelo fato de se ter um crescimento desorganizado (FREITAS *et al.*, 2014).

Ainda segundo a ONU (Organização das Nações Unidas), no século XXI, cerca de 2,8 bilhões de pessoas sofreram com desastres naturais, sendo que os danos excederam US$ 1,7 trilhão. Entre 1995 e 2015, as enchentes foram responsáveis por cerca de 26% das mortes por desastres naturais. No mesmo período, em todo o mundo, as inundações atingiram 56% das pessoas afetadas por algum tipo de desastre (SOUZA *et al.*, 2017).

Os problemas referentes às enchentes no Brasil vêm se agravando principalmente em períodos chuvosos (TUCCI; HESPANHOL; NETTO, 2003). Segundo (Souza *et al.*, 2018), existem mais de 40.000 áreas de riscos de enchentes que afetam mais de 120 milhões de pessoas que necessitam melhores preparo contra os efeitos das enchentes. Outro dado alarmante é o fato de que os riscos de enchentes estão ameaçando mais de 60% do PIB (Produto Interno Bruto) do Brasil.Esses gastos são referentes a bens materiais e inclusive às mortes provocadas por elas. Assim, existe uma busca enorme por estratégias com tecnologia de baixo custo para preparar as cidades e as pessoas contra os efeitos causados pelas enchentes.

No cenário de gerenciamento e monitoramento de desastres naturais, especialmente em

inundações em meios urbanos, tomar uma decisão rápida e assertiva é primordial, tendo em vista os riscos à vida e ao meio ambiente (YIN *et al.*, 2012).

Sendo assim, o ímpeto de unir tecnologias por meio de mineração é uma estratégia utilizada pelo que se denomina "cidades inteligentes" – que valem-se de dados obtidos por sensores físicos para prever esses desastres naturais e, assim, tornar as medidas mais assertivas para a solução desses problemas (LEMOS, 2013).

As Redes de Sensores Sem Fio (RSSF) surgem com uma tecnologia alternativa para aplicações de monitoramento. Basicamente, essa tecnologia é composta pelo nó *sink* (sorvedouro) e um conjunto de nós sensores sem fio e autônomos, com recursos energéticos limitados que podem ser móveis ou fixos (CARVALHO *et al.*, 2012). Dessa forma, esses sensores são dispostos aleatoriamente em um ambiente em mudança dinâmica - conhecido como campo de detecção, em que cada nó é um dispositivo de baixo consumo de energia. Contudo, existem algumas limitações na capacidade de sensoriamento desse sistema, atreladas às características particulares de cada ambiente monitorado (LOUREIRO *et al.*, 2003).

Na cidade de São Carlos - SP, Brasil, encontra-se instalada uma RSSF chamada e-noe, cujo objetivo é o monitoramento de rios a partir de sensores analógicos de pressão que aferem seus níveis de água. Esse projeto foi desenvolvido pelo Instituto de Ciências Matemáticas e de Computação (ICMC) da Universidade de São Paulo (USP) (PECHOTO; UEYAMA; PEREIRA, 2012).

Outra abordagem encontrada nas literaturas que discutem a identificação de enchentes são os modelos hidrológicos que, em geral, são definidos como uma representação matemática do fluxo de água e seus constituintes sobre alguma parte da superfície ou subsuperfície (SANTOS, 2009). Os modelos visam a simulação das previsões hidrológicas a partir de equações de caráter determinístico ou empírico (KEMP, 1993). Existem limitações básicas nesses modelos em relação à quantidade e qualidade de dados, além da complexidade em formular matematicamente o processo de calibração. Outro limite percebido em longo prazo é a necessidade de recalibração, devido ao fato de os dados não serem fixos, mas mudarem com o decorrer do tempo, gerando um alto custo para a manutenção desse modelo (TUCCI *et al.*, 1998).

Sendo assim, este trabalho tem como foco resolver os problemas existentes tanto em relação à complexidade e ao custo dos modelos hidrológicos quanto às limitações decorrentes da utilização de uma única fonte de dados, utilizada por outros modelos. Assim, é possível melhorar a identificação e resolução de enchentes. Este trabalho se propõe a desenvolver um modelo computacional para a identificação de enchentes, utilizando multivariáveis oriundas da mineração de dados de estações meteorológicas juntamente com os dados do projeto e-noe. Essa proposta tem como objetivo tornar o modelo computacional de classificação menos complexo e menos custoso, de maneira que este seja desenvolvido mais rapidamente sem perder a eficiência. Isso é possível por utilizar variáveis facilmente adquiridas, que não carecem do estudo do solo como os modelos hidrológicos, o que consome muito o tempo no processamento.

1.1 Motivação e Problema

A **motivação** desta dissertação relaciona-se ao contexto de enchentes – um tipo de desastre natural – e seus danos à sociedade. As enchentes são um grave problema nos grandes centros urbanos e ocorrem devido ao acúmulo da água das chuvas ocasionado pela inexistência de meios necessários para o seu escoamento. Dessa forma, a proposta é apresentar uma nova solução para que seja possível identificá-las e, assim, tornar possível um alerta para que a população e os órgãos responsáveis possam se prevenir de maneira mais eficiente, evitando danos maiores.

Para melhor exemplificar a motivação deste trabalho, demonstraremos como é o desenvolvimento de um modelo hidrológico. Serão apresentadas as camadas necessárias para a criação, tomando como base o trabalho de (FAVA *et al.*, 2018), realizado na mesma região do escopo desta dissertação. De maneira geral, o modelo necessita das três camadas explicadas a seguir:

- **Camada de Entrada:** onde são introduzidos os dados a serem analisados.
- **Camada de Modelagem:** entidade onde são desenvolvidas as equações matemáticas e físicas para calibrar o modelo. Por não se tratar de um modelo linear, essa parte é a mais crítica, pois é necessário grande volume de dados sobre o solo, dados físicos da bacia, batimetria e comprimento dos rios e tubulações. A calibragem pode ser feita por especialista experiente, tentativa e erro ou até algoritmo de otimização.
- **Camada de Saída:** a resposta do modelo, indicando possibilidade ou não de enchente.

Segundo Tucci *et al.* (1998), existem limitações nos modelos hidrológicos devido à **complexidade** e **custo** de desenvolvimento. A complexidade se dá na calibragem que é realizada na **camada de modelagem**. Basicamente, a calibragem, necessária para o desenvolvimento da camada, consiste em encontrar o padrão de manifestação das enchentes em determinada região. Esse problema é representado como um modelo não-linear.

Afirmar a **não-linearidade** na camada de modelagem significa que não existe uma maneira padrão para realização da calibração. Como foram estabelecidas várias maneiras diferentes para realizá-la, é difícil avaliar se o método utilizado obteve a melhor calibragem possível. Dessa maneira, desenvolver essa camada se torna uma tarefa não trivial.

Além disso, é igualmente importante observar que, com o passar do tempo, essa grande quantidade de dados necessários na camada de modelagem pode sofrer constantes mudanças. Assim, é imprescindível a atualização desses dados e, consequentemente, a **recalibração** do modelo. Essa constante atualização de dados resulta em um alto **custo** para a manutenção do modelo para que ele não perca a eficiência do seu funcionamento.

De modo geral, o modelo hidrológico demanda bastante tempo para ser desenvolvido. Isso é resultado da grande quantidade de dados e da dificuldade para a calibragem.

A partir disso, o **problema** está na complexidade,custo e na dificuldade de generalização dos modelos hidrológicos. Com isso, a proposta desta dissertação é apresentar um modelo multivariável que correlaciona dados do nível do rio e com as estações meteorológicas, adotando técnicas computacionais para a descoberta de padrões para identificar as enchentes. O resultado é a obtenção de um modelo menos complexo e custoso, com facilidade de generalização.

Portanto, esse modelo utiliza algoritmo de aprendizado de máquina para calibrá-lo, o que, consequentemente, diminui a sua complexidade. Outro ponto é que esse modelo tem a necessidade de uma única calibração, resultando na redução de custos. Ele é estruturado da seguinte forma:

- **Camada de Entrada:** camada onde são introduzidas as entradas do modelo, as variáveis dos dados de estações meteorológicas e dados do nível do rio.
- **Camada de Processamento:** entidade onde é utilizado o algoritmo de aprendizado de máquina para encontrar padrões e associá-los às enchentes.
- **Camada de Saída:** a resposta indicando se existe enchente ou não, baseada nos dados de entrada.

É importante ressaltar que todas as variáveis têm seu valor na confecção do modelo. No contexto do projeto e-noe, existem problemas como o assoreamento, as falhas e a necessidade de manutenções periódicas. Sendo assim, a criação do modelo apenas a partir dos dados do projeto e-noe não é recomendada, pois o funcionamento do sistema de alerta fica inteiramente dependente dos dados. O objetivo é o desenvolvimento de um modelo que possa identificar enchentes com autonomia dos dados do nível rio, utilizando esses dados apenas como referência para o aprendizado do modelo. Assim, cada dado tem sua relevância no contexto do modelo: os dados do nível do rio adquirido pelo projeto e-noe são utilizados apenas como um indicativo de enchente na fase de treinamento do modelo, enquanto os dados de estações meteorológicas são utilizados para a descoberta do padrão de enchentes.

Dessa forma, o modelo desenvolvido é construído relativamente mais rápido se comparado ao modelo hidrológico, por utilizar variáveis disponíveis mais facilmente, além do aprendizado de máquina, que encontra o padrão de enchentes nos dados e, assim, efetua a calibração do modelo.

1.2 Objetivos

O objetivo desta dissertação é desenvolver um modelo computacional de classificação que identifica enchentes e que utiliza multivariáveis, sendo os dados do nível do rio com os de estações meteorológicas que contém as seguintes variáveis: *(i)* Temperatura Máxima, *(ii)* Temperatura Miníma, *(iii)* Umidade, *(iv)* Precipitação e *(v)* Intensidade do vento, que foram indicadas

por especialistas. A metodologia utilizada foi a de mineração de dados do *Cross Industry Standard Process for Data Mining (CRISP-DM)* (BROWN, 2014), que permite mais objetividade para a criação do modelo. Esse modelo utiliza algoritmos de aprendizado de máquina para descobrir padrões nas bases de dados e, assim, determinar o indicativo de enchentes com menos complexidade. Ainda, o modelo necessita de uma única calibração na fase de treinamento, o que o diferencia em relação aos modelos hidrológicos da literatura. Esse modelo também emprega a ideia de usar diversas fontes de dados, não tornando a identificação dependente de uma única fonte de dados.

1.3 Estrutura do texto

Com as informações expostas acima, torna-se necessária a explanação de conteúdos teóricos que expandam a compreensão do texto. Em vista disso, distribuiremos os Capítulos da seguinte maneira: no Capítulo 2 discutiremos a fundamentação teórica em que a pesquisa será baseada; no Capítulo 3 ilustraremos a discussão com a exposição de trabalhos relacionados. No Capítulo 4 apresentaremos o modelo desenvolvido. Finalmente, a conclusão será apresentada no Capítulo 5.

CAPÍTULO

2

FUNDAMENTAÇÃO TEÓRICA

Esta seção tem o objetivo de introduzir os conceitos e as abordagens relacionados a este trabalho. A seção fica organizada assim : na Seção 2.1 são apresentados os conceitos relacionados à mineração de dados e sua principal contribuição no trabalho; na Seção 2.2 é introduzido o conceito de Aprendizado de Máquina, as técnicas e os algoritmos que serão utilizados no trabalho; na Seção 2.3 é apresentada a RSSF, seus conceitos e aplicação; na Seção 2.3.1 é descrito o e-noe, um sistema de RSSF instalado na cidade de São Carlos – SP para monitoramento de rios urbanos – trata-se de um estudo empírico do trabalho; na Seção 2.4 é descrito o conceito sobre os tipos de modelos hidrológicos e, finalmente, na Seção 2.5 é descrita a ferramenta de mineração de dados Weka.

2.1 Mineração de dados

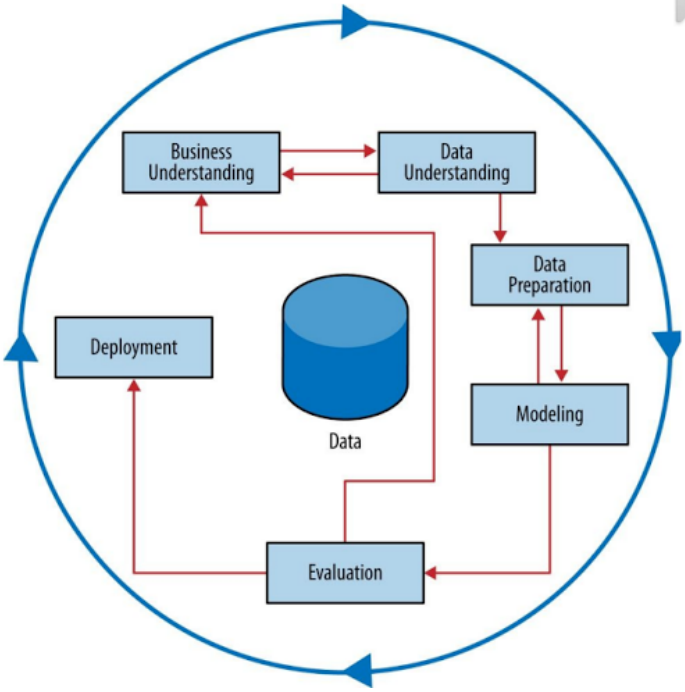
Mineração de Dados é a união de várias outras áreas, como tecnologias de bancos de dados e inteligências artificial e estatística. Trata-se de uma técnica de descoberta de informações que pode revelar conhecimentos de estruturas de dados que possam guiar decisões em condições de certeza limitada (AMO, 2004). Diversas definições de Mineração de Dados podem ser vistas na literatura. Destacamos as duas principais definições que exemplificam o conceito: i) Mineração de dados é a busca por informações de grande quantidade de banco de dados, é a cooperação entre o homem e o computador, sendo o homem o responsável por projetar os grandes bancos de dados, descrever os problemas e definir seus objetivos; enquanto o computador verifica os dados e procura padrões que se unam com as metas estabelecidas pelo homem, abstraindo informações relevantes (WEISS; INDURKHYA, 1998) e ii) Mineração de dados é explorar e analisar dados, de forma a obter padrões e adquirir conhecimentos interessantes e relevantes dos dados (LINOFF; BERRY, 2011).

No tocante à mineração de dados, a metodologia utilizada nesse trabalho é a *CRISP-DM* (*Cross Industry Standard Process for Data Mining*). Segundo (REIS *et al.*, 2017), a CRISP-

DM é um modelo de processo, que consiste em reunir boas práticas para Mineração de Dados. Além disso, esse processo embasa-se em ser repetitivo – pois permite retornar entre as camadas, caso seja necessário – e objetivo – por conseguir agir diretamente sobre o problema – conforme a necessidade do projeto. Dessa forma, esse modelo é constituído por 6 etapas que são, conforme a Figura 1:

- **Entendimento do negócio**: a etapa responsável por analisar os objetivos do projeto, levando em consideração os problemas da aplicação.
- **Entendimento dos dados**: resolver o problema é o principal objetivo. A compreensão dos dados é a matéria prima para que a solução seja construída. Dessa forma, o objetivo é conhecer características e limitações das bases de dados, o histórico, sua composição, seu tipo e se os dados realmente são aceitáveis para resolver o problema proposto.
- **Preparação dos dados**: é a tarefa mais árdua, pois envolve todas as atividades associadas à construção do conjunto final de dados, aquele que será usado na ferramenta de modelagem, sofrendo inevitavelmente várias otimizações. Nesse processo, é necessário aplicar quatro tarefas:
  - **Seleção dos Dados:** nessa parte são selecionados os dados que serão utilizados no modelo. Todos os dados essenciais são escolhidos para alcançar o objetivo proposto.
  - **Limpeza dos Dados:** etapa utilizada para realizar a limpeza dos dados, como, por exemplo, datas em formato incorreto e números inteiros sendo interpretados como *strings* e etc.
  - **Construção dos Dados:** construir uma nova base de dados, algo que seja relevante para a base de dados em um todo.
  - **Integração dos Dados:** junção de duas fontes de dados diferentes.
- **Modelagem**: é o momento em que serão utilizadas técnicas mais aderentes ao objetivo do projeto, seja ele uma predição, classificação, agrupamento ou regressão. A etapa de modelagem pode manter um canal de comunicação contínuo com a etapa de preparação de dados, seja para a readequação dos dados ou mesmo para a criação de novas variáveis que ajudem a explicar o fenômeno. Nessa etapa, pode-se criar diferentes modelos e compará-los na próxima etapa.
- **Avaliação**: tem como finalidade avaliar a utilidade do modelo, rever passos executados na sua construção e verificar se permitem atingir os objetivos;
- **Implementação**: nesta etapa, o modelo é encerrado com a entrega. É a etapa menos técnica do processo de Mineração de Dados, mas não a menos importante.

Figura 1 – Fases da metodologia CRISP-DM



Fonte: Reis *et al.* (2017)

A premissa de Mineração de Dados é uma **argumentação ativa**, isto é, ao invés do usuário definir o problema, selecionar os dados e as ferramentas para analisar tais dados, as ferramentas de mineração de dados pesquisam automaticamente à procura de anomalias e possíveis relacionamentos. Assim, identificam problemas que não tinham sido identificados anteriormente, tornando a aplicação mais eficiente.

A Mineração de Dados é uma área que está se consolidando cada vez mais pois, com a geração de um volume de dados cada vez maior, é essencial utilizar técnicas para retirar a maior quantidade de informações possível desses dados. Talvez a forma mais nobre de se utilizar esses vastos repositórios seja tentar descobrir se há algum conhecimento escondido neles.

2.2 Aprendizado de Máquina

O aprendizado de máquina – uma subárea da Inteligência Artificial –, está preocupado em determinar padrões e alterar seus comportamentos de forma adequada para resolver com maior precisão um determinado problema. Logo, o ponto principal é ter a capacidade de generalizar um determinado comportamento a partir de um novo cenário, melhorando automaticamente a partir das próprias experiências. Em aprendizado de máquina, existem várias abordagens possíveis para cada situação, tais como: I) Aprendizagem supervisionada; II) Aprendizagem não-supervisionada e III) Aprendizagem por esforço (MITCHELL, 1997). Entretanto, este trabalho

utiliza a aprendizagem supervisionada.

No aprendizado de máquina supervisionado, o algoritmo utiliza um modelo de treinamento, onde cada x está ligado a um y. Dessa forma, o algoritmo descobre o padrão para que possa, futuramente, ser capaz de predizer o valor de y para um novo x, isto é, o valor de uma função ou a classe à qual um novo exemplo pertence. Dentro desse contexto, existe a técnica de classificação ou regressão. A saída que cada uma oferece é o que as difere. Caso a saída de y seja **contínua**, em que não existe um valor fixo e sim o valor aproximado, dizemos que é uma **técnica de regressão**. Por sua vez, caso a saída de y seja um valor **discreto** – no conceito de ser um valor fixo (por exemplo, 1 ou 0) – dizemos que é uma **técnica de classificação**. A escolha da melhor técnica está vinculada ao tipo de problema a ser abordado. Assim, nesse trabalho, a técnica utilizada é a de classificação. Nas próximas subseções, o conceito e o funcionamento de cada algoritmo escolhido no ambiente de desenvolvimento deste trabalho serão explicados mais pormenorizadamente.

2.2.1 Naive Bayes

Naive Bayes é uma técnica estatística baseada no teorema de Thomas Bayes, seguindo o paradigma de probabilidade. Esse teorema descreve a probabilidade de um evento, baseado em um conhecimento a *priori* que pode estar relacionado a ele. Desse modo, mostra como alterar as probabilidades a *priori* tendo em vista novas evidências que podem surgir posteriormente (EHLERS, 2007).

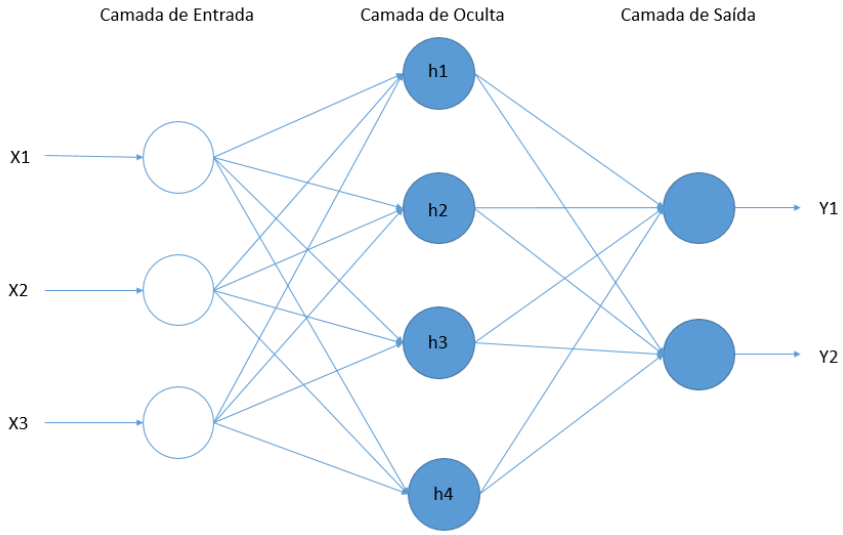
Num contexto geral, o algoritmo de *Naive Bayes* define que não existe uma certa dependência entre os atributos, assumindo que o valor de uma característica particular é independente do valor de qualquer outra característica, dada a variável de classe. Por exemplo, uma fruta pode ser considerada um melão se for verde, redondo e com cerca de 20 cm de diâmetro. Um algoritmo de *Naive bayes* considera que cada uma dessas características contribui independentemente para a probabilidade de essa fruta ser um melão, independentemente de quaisquer possíveis correlações entre as características de cor, redondeza e diâmetro.

2.2.2 Multilayer Perceptron (MLP)

O algoritmo de *Multilayer Perceptron* (MLP) é uma classe de *rede neural artificial feedforward*. Esse algoritmo é baseado em um paradigma de função e sua estrutura é basicamente formada por três camadas de nós: uma camada de entrada, uma ou mais camadas ocultas e uma camada de saída, como observado na Figura 2. Cada camada é um sistema de neurônios interconectados, em que os nós são conectados por pesos e sinais de saída, ocorrendo uma função da soma das entradas para o nó modificado por uma simples função de transferência ou ativação não-linear. A superposição de muitas funções simples de transferência não-linear é que permite ao algoritmo se aproximar de funções extremamente não-lineares (GARDNER; DORLING,

1998a). Esse algoritmo permite realizar tanto a classificação ou regressão, oriundas de técnicas supervisionadas. O treinamento supervisionado do algoritmo MLP consiste em dois passos. No primeiro, um padrão é apresentado às unidades da camada de entrada e, a partir desta camada, as unidades calculam sua resposta que, por sua vez, é produzida na camada de saída. O erro é calculado e, no segundo passo, ele é propagado a partir da camada de saída até a camada de entrada. Os pesos das conexões das unidades das camadas internas vão sendo modificados utilizando a regra delta generalizada. Este processo é repetido até atingir algum critério de parada, o que acarreta a diminuição do erro (BRAGA, 2007).

Figura 2 – Exemplo de Multilayer Perceptron



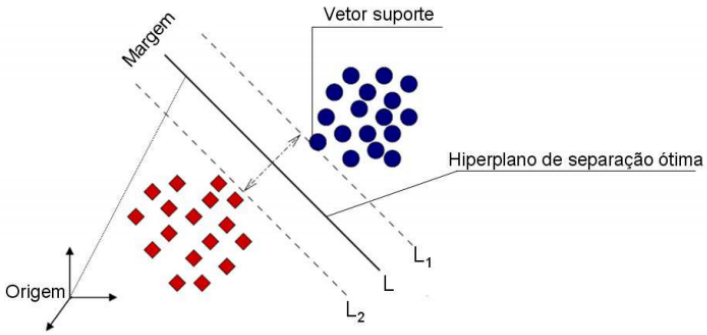
Fonte: Gardner e Dorling (1998b)

2.2.3 Support Vector Machines (SVM)

Support Vector Machines (do português, Máquina de Vetores Suporte) é um algoritmo de aprendizado de máquina supervisionado que pode ser utilizado tanto para classificação quanto para regressão, empregado no paradigma de função. As SVMs são embasadas pela teoria de aprendizado estatístico. Essa teoria estabelece uma série de regras que devem ser obedecidas para que os classificadores obtenham uma boa generalização, definida como a sua capacidade de maior acerto e consequentemente menor erro das classes dos novos dados que foram aprendidos, definido por (VAPNIK, 2013), o criador dessa teoria. Como apresentado na Figura 3, o algoritmo busca a construção de um hiperplano(L) orientado para maximizar a margem (distância entre as bordas, L1 e L2) de forma que a separação das classes seja bem mais perceptiva. Existem dois tipos de SVMs: linear e não-linear. Segundo (LORENA; CARVALHO, 2007), os SVMs lineares são bons para conjuntos de dados linearmente separáveis. Em contrapartida, para dados que não podem ser separados por um hiperplano, temos o SVM não-linear. Para casos do uso de dados não separados linearmente, uma possibilidade é a utilização da técnica de *kernels* que aumenta a

dimensionalidade do espaço amostral, permitindo que os dados possam ser separados por um hiperplano.

Figura 3 – Exemplo do SVM



Fonte: Nascimento *et al.* (2009)

2.2.4 K-Nearest Neighboors (KNN)

Esse algoritmo é um dos classificadores mais tradicionais. O *K Nearest Neighboors* (KNN) – em tradução livre, K Vizinhos Mais Próximos –, emprega a técnica de paradigma de busca. Ele foi proposto por (FUKUNAGA; NARENDRA, 1975). A ideia principal desse algoritmo é determinar o rótulo de classificação de uma amostra de dados baseado no comportamento de que dados próximo tendem a estar em uma mesma classe. Sendo assim, um grupo de técnicas denominado *Instance-based Learning* é utilizado como base, onde são encontrados os k vizinhos mais próximos dos dados de modelo de treinamento.

As duas principais características desse algoritmo são: a regra de classificação e a função que calcula a distância entre duas instâncias. A primeira determina como o algoritmo vai tratar a importância de cada um dos k elementos selecionados – os k mais próximos. E, na função de distância, cabe a tarefa de mensurar a diferença entre dois elementos para que se possa identificar quais são os vizinhos mais próximos. São esses os parâmetros que podem ser alterados conforme a necessidade exigida pelo problema.

2.2.5 Random Forest

A Floresta Aleatória (do inglês, *Random Forest*), é um algoritmo de aprendizado supervisionado que constrói uma multiplicidade de árvores de decisões (HO, 1995). De maneira simplificada, ela é assim denominada porque funciona como um fluxograma em forma de árvore, em que cada nó indica um teste realizado, baseado em uma condição específica. As ligações entre os nós representam os valores possíveis do teste do nó superior, e as folhas indicam a classe à qual o dado pertence. Esse algoritmo utiliza a estratégia de dividir para conquistar: um problema complexo é decomposto em subproblemas mais simples e, recursivamente, esta técnica

é aplicada a cada subproblema. (CREPALDI *et al.*, 2011). Dessa forma, a Floresta Aleatória é uma coletânea de árvores de decisão e o método utilizado no treinamento é o método de ensacamento(do inglês, *Bagging*), que, de maneira geral, consiste na combinação de uma pluralidade de modelos de aprendizagem gerados a partir das árvores como base para uma boa generalização, aumentando a performance do modelo. Uma das diferenças entre a floresta aleatória e árvore de decisão é que a floresta aleatória é facilmente generalizada e impede o *overfitting* [1], pois cria subconjuntos aleatórios dos recursos, construindo árvores menores.

2.3 Redes de Sensores Sem Fio - RSSF

Segundo (LOUREIRO *et al.*, 2003), RSSFs podem ser vistas como um tipo especial de rede móvel *ad hoc* intitulado de *Mobile Ad hoc Network* (*MANET*). Assim as *MANETs* têm como função básica prover um suporte à comunicação entre esses elementos computacionais que, individualmente, podem estar executando tarefas distintas. Por outro lado, RSSFs tendem a executar uma função colaborativa onde os elementos (sensores) proveem dados que, adiante, serão processados (ou consumidos) por nós especiais chamados de sorvedouros (nó *sink*).

Já (AKYILDIZ *et al.*, 2002) descreve a RSSFs como sistemas cooperativos de baixo custo e de baixo consumo de energia que, normalmente, utilizam comunicação *multihop*, sendo independentes de serviços externos. Essas redes são constituídas por nós que, por sua vez, consistem em uma interface sem fio, dispositivos de sensoriamento, bateria, processador e memória. As RSSFs são compostas por nós sensores e um nó *sink*, que podem estar dispostos perto ou dentro dos fenômenos a serem monitorados. Na Figura 4 serão apresentados os dois dispositivos que compõem essa arquitetura:

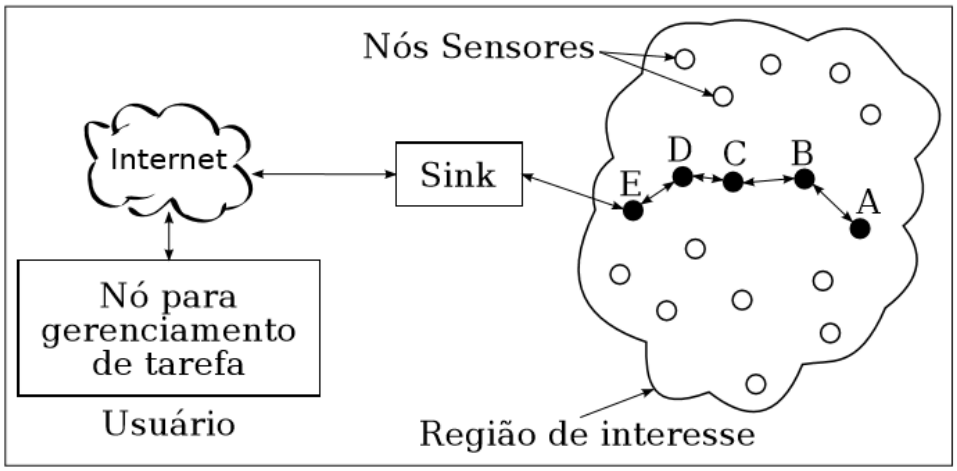
- **Nós Sensores:** menor capacidade computacional, alocados em grande quantidade dentro da região que se deseja monitorar.

- **Nó *Sink*:** maior capacidade computacional cuja função primária é realizar a ligação entre as RSSF e o mundo exterior. Seria um nó para gerenciamento de tarefa, que normalmente utiliza protocolos nativos da internet.

Cada nó sensor é capaz de executar algumas funções, tais quais:(*i*) adquirir dados por meio de sensoriamento e processá-los, podendo detectar eventos, (*ii*) comunicar esses eventos a seus vizinhos por meio de um rádio e (*iii*) auxiliar na propagação de dados/informações de seus vizinhos para que o evento seja entregue a um Nó *Sink*. Essas funções são essenciais em uma RSSF, tornando-se imprescindíveis para definir ações que devem ser tomadas no ambiente monitorado.

[1] *Overfitting* é um termo usado em estatística para descrever quando um modelo se ajusta muito bem ao conjunto de dados anteriormente observado, mas se mostra ineficaz para prever novos resultados.

Figura 4 – Arquitetura da RSSF



Fonte: Traduzido de Akyildiz *et al.* (2002)

Convém salientar que um nó sensor é composto por dois componentes-chaves: *i) hardware de interface de comunicação* e *ii) Conjunto de sensores*, como descrito no trabalho (AKYILDIZ *et al.*, 2002). Assim, os componentes e suas funções são dispostos da seguinte maneira:

- ***Hardware* de Interface de comunicação**: é o principal componente dos nós sensores por possuir capacidade de comunicação e programação. Habitualmente é composto por um microcontrolador, um rádio para comunicação, fonte de alimentação e memória. Algumas interfaces podem dispor de sensores, como sensor de temperatura, pressão, umidade, luminosidade, entre outros;

- **Conjunto de sensores**: Utilizados para capturar dados do ambiente em que estão inseridos e podem ser acoplados aos Hardware de interface com o intuito de aumentar suas capacidades sensitivas;

Portanto, as RSSFs tornaram-se soluções atrativas para serem utilizadas em diversas aplicações de monitoria de diferentes fenômenos no ambiente. Entretanto, notou-se que cada aplicação possui sua peculiaridade e, por isso, devem ser implantadas de maneira eficiente, considerando todas as suas limitações.

2.3.1 Projeto e-noe

E-noe (PECHOTO; UEYAMA; PEREIRA, 2012) é um projeto brasileiro de monitoramento de rios urbanos liderado pelo Instituto de Ciência Matemática e de Computação (ICMC) da USP que conta com a parceria de hidrólogos da Escola de Engenharia de São Carlos (EESC), também da USP. O propósito desse projeto é prover condições de monitoramento de rios urbanos contra a poluição e principalmente enchentes, a fim de gerar alertas e evitar danos maiores, como perdas de vidas. Esse projeto foi instalado no município de São Carlos - SP sendo composto por

nós sensores e nó *sink* espalhados pelo leito do rio. Para sua realização, os nós sensores estão equipados com sensor de pressão. O objetivo desse sensor é obter com facilidade a informação da altura de água do rio monitorado. Sendo assim, se essa altura variar rapidamente dentro de um curto intervalo de tempo, qualifica-se como um sinal de enchente.

Figura 5 – Disposição dos Nós do Projeto e-noe

Fonte: PECHOTO, UEYAMA e Pereira (2012)

A Figura 5 mostra os lugares onde cada nó está instalado em São Carlos-SP. Esses pontos foram previamente estudados por causarem enchentes:

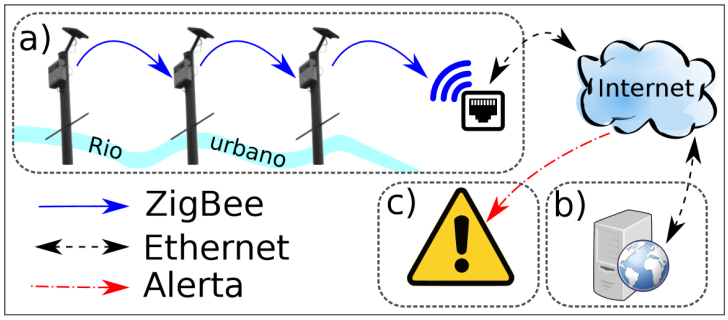
- Ponto 1: Ponto na USP Base 1;
- Ponto 2: Ponto do kartódromo 1;
- Ponto 3: Ponto do kartódromo 2;
- Ponto 4: Ponto da USP;
- Ponto 5: Ponto Sesc;
- Ponto 6: Ponto do Cristo;

Já a Figura 6 apresenta o funcionamento do projeto e-noe (VIEIRA, 2015), fundamentado em 3 etapas principais, sendo elas: a) Transmissão dos dados coletados dos rios pelos Nós sensores até o Nó *sink* (Estação Base); b) Armazenamento de Dados e predição de inundações e c) Geração de Alertas para a população ribeirinha;

2.4 Modelos Hidrológicos

Os modelos hidrológicos procuram representar a parte terrestre do ciclo hidrológico. Eles transformaram a precipitação que cai sobre a bacia em vazão numa determinada seção de um rio (ALMEIDA; SERRA, 2017). Segundo (TUCCI; HESPANHOL; NETTO, 2003), o

Figura 6 – Arquitetura do Projeto e-noe



Fonte: Vieira (2015)

desenvolvimento do modelo consiste em uma representação simplificada do sistema do mundo real com equações de caráter determinístico ou empírico. De modo geral, os modelos hidrológicos são classificados, dentre outras formas, de acordo com os tipos de variáveis utilizadas na modelagem, os tipos de relações entre essas variáveis, a forma de representação dos dados, a existência ou não de relações espaciais e a existência de dependência temporal (ALMEIDA; SERRA, 2017). Assim, (DEVIA; GANASRI; DWARAKISH, 2015) demonstram três tipos:

- **Modelo Empírico:** necessita de informações de dados do ambiente monitorado sem carecer de informações sobre as suas características e os processos do ambiente físico. Envolve equações matemáticas derivadas de séries temporais de entrada e saída concorrentes, não de processos físicos da bacia. A calibração ocorre por meio de técnicas de inteligência artificial atreladas à correlação dos dados;

- **Modelo Conceitual:** baseado em equações que descrevem o processo físico conceitual ou hipotético, não sendo necessariamente baseado no processo real. Equações semi-empíricas são usadas nesse método e os parâmetros do modelo são avaliados não apenas a partir de dados de campo, mas também através de calibração. A calibração envolve ajuste de curva que dificulta a interpretação. Sendo assim, o efeito da mudança que ocorre no ambiente não pode ser previsto com muita confiança. É um modelo complexo no contexto da calibração, um fator importante ligado na exatidão do modelo;

- **Modelo Físico:** representação matematicamente idealizada do fenômeno real. Também chamado de modelo mecânico porque inclui os princípios dos processos físicos. Utiliza variáveis de estado mensuráveis e são funções de tempo e espaço. Por ser considerado um modelo não-linear, os processos hidrológicos do movimento da água são representados pela diferença infinita de equações. É necessário um grande número de parâmetros que descrevam as características físicas da captação, tornando-o complexo na calibração e ao mesmo tempo custoso em longo prazo, por possuir parâmetros intermitentes que afetam sua eficiência.

Assim, o modelo desenvolvido se encaixa na categoria empírica. Esse modelo é considerado menos custoso e complexo em comparação aos outros por possuir um desenvolvimento mais simplificado para a calibração, além de não perder a eficiência em longo prazo. Ademais, a sua facilidade de generalização é uma de suas virtudes para a identificação de enchentes.

2.5 Weka

A ferramenta Weka é um framework do tipo *open source* (software livre) de mineração de dados, desenvolvido em Java dentro das especificações da *GPL (General Public License)* que, por sua vez, se consolidou como a ferramenta de mineração de dados. Ela foi desenvolvida pela Universidade de Waikato na Nova Zelândia e tem como objetivo agregar algoritmos provenientes de diferentes paradigmas na sub-área de Inteligência Artificial dedicada ao estudo de aprendizagem de máquina. Uma das suas principais características é a portabilidade, pois, além de aproveitar os principais benefícios da orientação a objetos, é possível utilizá-la em diferentes sistemas operacionais. (HALL *et al.*, 2009). Com a ferramenta Weka é possível utilizar diversas técnicas, dependendo do objetivo do projeto. Por meio do trabalho de (KOTSIANTIS; ZAHARAKIS; PINTELAS, 2007) é feito um resumo das técnicas disponíveis na ferramenta, sendo elas:

- **Técnicas de Classificação:** o algoritmo de aprendizado recebe uma quantidade de exemplos de treinamento, a partir dos quais o rótulo da classe associada é conhecido. Dessa maneira, cada exemplo é descrito por um vetor de valores de características ou atributos, levando em conta o rótulo da classe associada. O objetivo do algoritmo de indução é ser capaz de construir um classificador que possa determinar corretamente a classe de novos exemplos ainda não vistos;

- **Técnica de Regressão:** o algoritmo recebe uma base de dados que não possui um modelo de exemplo para a classificação. Parte da tarefa do algoritmo é prever futuras situações, baseado nos estudos dos dados. A partir de uma entrada de dados é que as saídas previstas para aquele quadro são construídas. Neste caso, o algoritmo deve se ajustar para chegar aos resultados corretos e com o máximo de acertos. Para isso, o aprendizado pode ser constante, aumentando, assim, a experiência com aquele problema;

- **Técnica de Clusterização:** o algoritmo analisa os exemplos fornecidos e tenta determinar se alguns deles podem ser agrupados de alguma maneira, podendo descobrir relações, padrões, regularidades ou categorias nos dados que lhe vão sendo apresentados. Não existe um conjunto de modelo específico, pois o próprio algoritmo encontra separações possíveis a partir dos dados;

- **Técnica de Associação:** a técnica de Associação consiste na identificação de padrões

intrínsecos ao conjunto de dados, ou seja, encontra conjuntos de itens que ocorram simultaneamente e de forma frequente em um banco de dados.

Dessa forma, existem vários algoritmos dispostos em cada técnica, tais quais são:

- **Técnica de Classificação:** Árvore de decisão, Regras de aprendizagem, *Naive Bayes*, Tabelas de decisão, Regressão local de pesos, Aprendizado baseado em instância, Regressão lógica, Perceptron, Perceptron multicamada, Comitê de perceptrons, *Support Vector Machines* (SVM), *K-Nearest Neighboor* (KNN) e Floresta Aleatória;
- **Técnica de Regressão:** Regressão linear, Geradores de árvores modelo, Regressão local de pesos, Aprendizado baseado em instância, Tabela de decisão e Perceptron multicamadas;
- **Técnica de Clusterização:** *EM*, *Cobweb*, *SimpleKMeans*, *DBScan* e *CLOPE*;
- **Técnica de Associação:** *Apriori*, *FPGrowth*, *PredictiveApriori*, *Tertius*;

CAPÍTULO

3

TRABALHOS RELACIONADOS

Neste capítulo, serão apresentados alguns trabalhos na área de computação e de hidrologia que abordam a realização dos sistemas baseados em alertas, identificação e arquiteturas relacionados a desastres naturais no caso de enchentes.

(FURQUIM *et al.*, 2018) apresenta uma arquitetura intitulada SENDI (Sistema de Detecção e Previsão de Desastres Naturais baseado em *IoT*). SENDI é um sistema tolerante a falhas baseado em *IoT*, aprendizado de máquina e RSSF para a detecção e previsão de desastres naturais e para a emissão de alertas. A tolerância a falhas está embutida no sistema, antecipando o risco de falhas na comunicação e destruição dos nós durante os desastres. Por utilizar uma única variável, o sistema depende integralmente dela para identificar as enchentes. Sua acurácia também é limitada por utilizar uma única variável.

(ACOSTA-COLL; BALLESTER-MERELO; MARTÍNEZ-PEIRÓ, 2018) descrevem um sistema de alerta precoce de baixo custo para detectar, em tempo real, o nível de risco de um fluxo em uma bacia não utilizada. O sistema indica se é seguro ou não atravessar a rua alagada. Um modelo hidrológico e hidráulico calcula o fluxo, a velocidade e o nível de água em todas as seções transversais ao longo do fluxo. O modelo usa apenas medições em tempo real de medidores de chuva e dados de levantamentos topográficos para determinar o nível do perigo. No entanto, nesse trabalho o problema reside nas informações topográficas para realizar a identificação de enchentes. A calibração acaba sendo custosa e complexa. Além disso, a grande quantidade de dados e as constantes mudanças que eles podem sofrer com o tempo acabam por influenciar a eficiência do modelo, que necessita de recalibração.

(MOSTAFA; MOHAMED *et al.*, 2014) apresentam um modelo inteligente que reúne dados recebidos dos sensores sem fio e os alcança de forma inteligente. Por um lado, detecta dados errôneos ou redundantes para apresentar apenas os dados confiáveis e adequados. Os dados armazenados no banco são processados no sistema dando suporte à decisão para a previsão de inundação em tempo real, conciliando o sistema multi-agente (MAS) para o processamento dos

dados, eliminar dados redundantes não úteis e estabelecer a colaboração entre agentes móveis para enviar os resultados para a estação base. Essa abordagem utiliza diversas fontes a partir de agentes inteligentes para tomar a decisão sobre as enchentes. Nesse trabalho não foi apresentada a forma de calibração feita no sistema, nem tampouco informa se dados históricos são utilizados para realizar a identificação de enchentes, além de não demostrar se os dados foram coletados em um ambiente real.

(CHEN *et al.*, 2013) descreve um sistema de alerta para riscos geológicos na região do reservatório, que se baseia na tecnologia RSSF. O trabalho tem como foco questões como: *(i)* suporte à transmissão de dados confiáveis, *(ii)* manipulação de dados enormes de tipos e fontes heterogêneas e tipos e *(iii)* minimizar o consumo de energia. Este estudo propõe um protocolo de roteamento dinâmico, um método para a recuperação de rede e um método para gerenciar nós móveis para permitir a transmissão de dados em tempo real e confiável. O sistema incorpora abordagens de fusão e reconstrução de dados para reunir todos os dados em uma única visão do risco geológico sob monitoramento. Esse trabalho não informa qual técnica foi utilizada para gerar o alerta sobre o risco de inundação nessas regiões próximo do reservatório. Esse é mais baseado na arquitetura do que na geração de um modelo.

(BOTH *et al.*, 2008) tem como objetivo o desenvolvimento de um modelo de previsão de enchentes na região do Vale do Rio Taquari, situada na Bacia Hidrográfica Taquari-Antas, no Rio Grande do Sul. O método de previsão foi desenvolvido a partir de dados adquiridos de vários órgãos e entidades regionais e estaduais. Assim, um equação matemática foi elaborada para correlacionar esses dados e desenvolver o modelo. Uns dos problemas desse trabalho é a dificuldade de generalização, pelo fato de não ter utilizado nenhum RSSF, nem consolidar a possibilidade de geração de alertas.

(FAVA, 2015) apresenta uma nova proposta metodológica de previsão de enchentes: o Modelo de Alerta Hidrológico com Base Participativa (MAHP). Trata-se de um modelo de previsão de enchentes em bacias urbanas que integra as Informações Geográficas Voluntárias (VGI) e as redes de sensores sem fio. O modelo MAHP foi dividido em módulos, sendo que cada um deles é responsável por uma atividade no processo de previsão de enchentes. Embora possua diversos módulos auxiliares, pode-se resumir o modelo MAHP em três módulos principais: aquisição de dados; calibração por fórmulas físicas a partir dos dados topológicos; e, por fim, o módulo responsável pela previsão das enchentes. Dessa forma, os problemas desse trabalho estão relacionados à complexidade na calibração e ao custo em longo prazo do mesmo. Como esse modelo precisa de uma grande quantidade de dados e esses dados estão propícios a sofrerem alterações com o passar do tempo, serão sempre necessárias atualizações das variáveis de calibração e, consequentemente, recalibrações.

3.1 Discussão dos Trabalhos Relacionados

Após o levantamento de trabalhos computacionais e hidrológicos, pode-se perceber que só alguns utilizam o conceito de multivariáveis para detectar enchentes.

Já nos modelos hidrológicos, nota-se a complexidade e o custo, tanto para o seu desenvolvimento, quanto para mantê-los funcionando com precisão em longo prazo. Isto porque as variáveis sofrem alterações com o passar do tempo, necessitando de frequentes recalibrações para continuarem funcionando.

Assim, o modelo proposto neste trabalho utiliza a mineração de dados em multivariáveis, ou seja, valendo-se de diversas fontes de dados. A calibração é realizada por técnicas de aprendizado de máquina que descobrem padrões nos dados sobre o evento de enchente, diminuindo a complexidade e o custo na calibragem em relação aos trabalhos da hidrologia. E também, resolvendo a questão da dependência de uma única variável de alguns modelos computacionais, a partir do momento que utiliza multivariáveis. A Tabela 1 apresenta um resumo das principais características dos trabalhos:

Tabela 1 – Comparativo das principais características dos trabalhos relacionados

Trabalhos Analisados	RSSF	Alerta	I.A	Multi-Fontes	Ambiente Real
(CHEN *et al.*, 2013)	**X**	**X**	**-**	**-**	**X**
(ACOSTA-COLL; BALLESTER-MERELO; MARTÍNEZ-PEIRÓ, 2018)		**X**	**X**		**X**
(FURQUIM *et al.*, 2018)	**X**	**X**	**X**		**X**
(MOSTAFA; MOHAMED *et al.*, 2014)		**X**		**X**	**-**
(FAVA, 2015)	**X**	**X**			**X**
(BOTH *et al.*, 2008)			**X**	**X**	**X**
Trabalho de Mestrado	**X**	**X**	**X**	**X**	**X**

CAPÍTULO

4

MODELO DE CLASSIFICAÇÃO MULTIVARIÁVEL PARA IDENTIFICAÇÃO DE ENCHENTES

4.1 Considerações iniciais

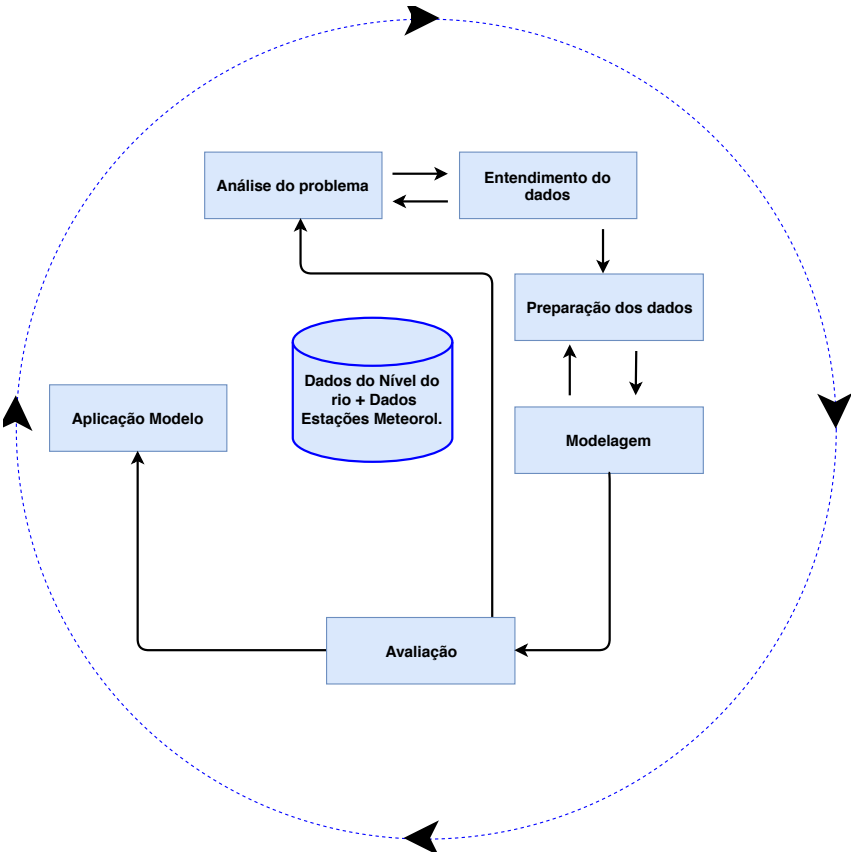
O foco principal do modelo é desenvolver um mecanismo para a identificação de enchentes. Para alcançá-lo é realizada a mineração de dados em diversas fontes, correlacionando dados de estação meteorológica com os do nível do rio. Para adquirir os dados do nível do rio, o trabalho realiza um estudo empírico no projeto e-noe. Assim, o modelo desenvolvido pode ajudar a mitigar os problemas que as enchentes podem provocar. Este capítulo é organizado da seguinte maneira: a seção 4.2 apresenta o Modelo de identificação de enchentes, com a explicação da metodologia utilizada e de todas as decisões no desenvolvimento do modelo; a seção 4.3 expõe os resultados que demostram a forma de avaliação para a escolha do melhor algoritmo para o modelo.

4.2 Modelo de Identificação de Enchentes

Este trabalho foi desenvolvido utilizando um modelo de classificação com multivariáveis de identificação de enchentes, a partir do uso da técnica de mineração de dados. Para isso, recorre às bases de estações meteorológicas e do nível do rio, adquiridas pelo Projeto e-noe. Dessa forma, a metodologia do *CRISP-DM* (*Cross Industry Standard Process for Data Mining*) foi implementada como descrita no (BROWN, 2014).Trata-se de um modelo de processo para a realização de mineração de dados. Esse modelo reúne as melhores práticas para que a técnica de mineração de dados seja usada da forma mais produtiva possível (KURGAN; MUSILEK, 2006). Além disto, apesar de ser composto por fases, ele é flexível e interativo entre as etapas.

Essa característica permite aprimorar o desenvolvimento do modelo conforme a necessidade do projeto. A Figura 7 demostra a composição de todas as etapas do *CRISP-DM*. As próximas subseções explicam a importância de cada etapa no contexto da proposta.

Figura 7 – Fases da metodologia CRISP-DM



Fonte: Adaptado de Reis *et al.* (2017)

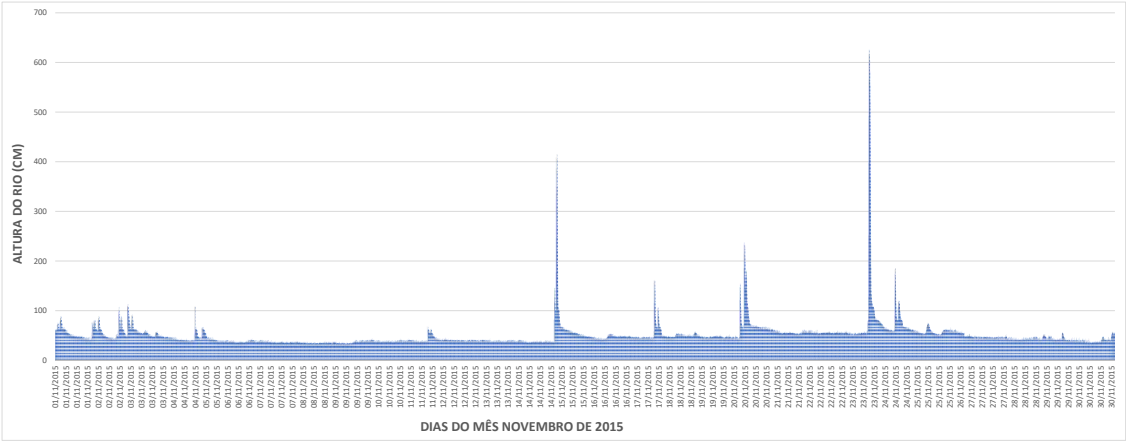
4.2.1 Análise do Problema

Essa etapa é responsável pela definição do objetivo da proposta, levando em consideração os problemas da aplicação para o desenvolvimento do modelo. O ponto escolhido para extrair os dados e compor a base é o **ponto** 6 **- Ponto do Cristo** do projeto e-noe, por ser o ponto de intersecção entre as bacias que se encontram em São Carlos-SP com menor altitude, o que ocasiona maior quantidade de enchentes.

Para exemplificar o problema das enchentes, resolvemos demonstrar um estudo realizado por (BRITO; BRESSIANI; UEYAMA, 2018), que apresentam o comportamento do rio no mês de novembro de 2015. Nesse mês, o rio atingiu seu maior ápice, ultrapassando 600 centímetros - seis metros - de altura, conforme a leitura do sensor. Nesse ponto 6 - Ponto do Cristo, quando o rio está acima de 200 centímetros - dois metros -, considera-se o risco de enchentes. Esse

valor foi determinado por um estudo nesse mesma região da nossa análise relatado no trabalho de (FURQUIM, 2017). A Figura 8 demostra em detalhes como a altura do nível do rio varia diariamente. Vale ressaltar que a média do mês de novembro chegou a 93 centímetros, uma média considerada alta em comparação aos outros meses dos outros anos.

Figura 8 – Gráfico da variação do nível do rio no mês de novembro de 2015



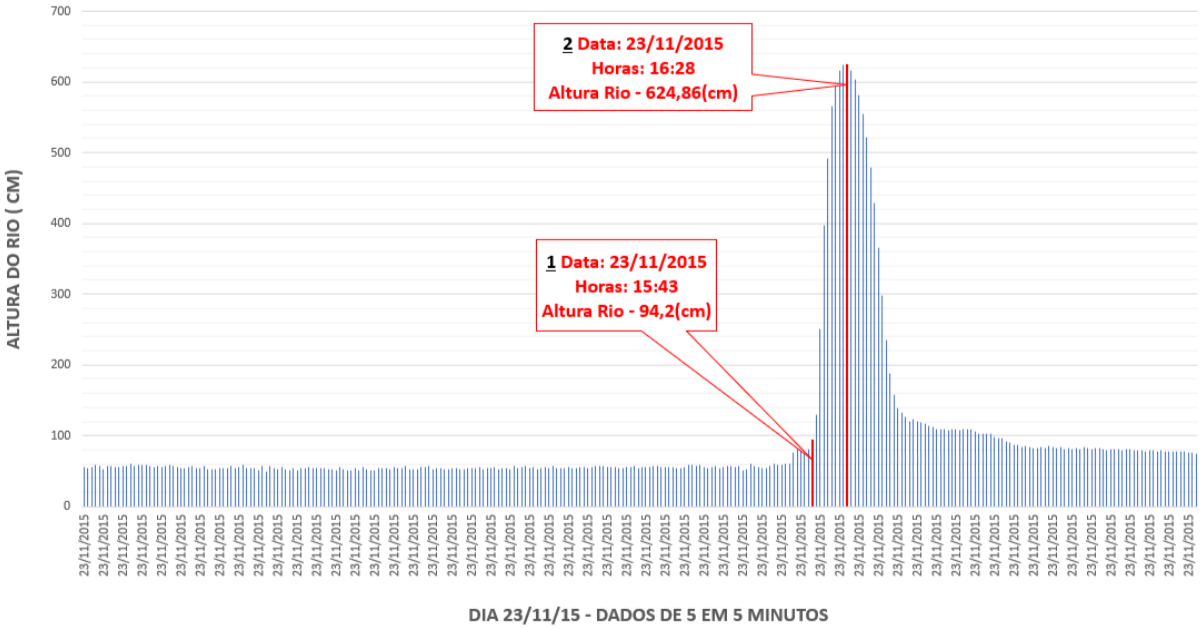
Fonte:gráfico elaborado pelo autor.

Com o propósito de demonstrar a altura que o rio pode alcançar, foi criado um gráfico apenas com o dia 23/11/15, data em que o rio alcançou sua maior altura no mês de novembro.

A Figura 9 apresenta a velocidade do aumento do nível do rio. Conforme analisamos, a **origem** 1 demonstra o nível às 15h43, em que atinge $94,2$ centímetros de altura, enquanto a **origem** 2, pouco mais de 45 minutos depois, ele chega a $624,86$ centímetros de altura, representando assim um aumento de 665% em um curto espaço de tempo, tornando esse desequilíbrio interessante para o estudo.

Após essas análises, é demostrado como o desenvolvimento de mecanismos para a identificação de enchentes é importante. A utilização de uma única variável pode tornar o modelo dependente. Imaginando um cenário de falha desse dado do nível do rio, a identificação não seria mais possível. Sendo assim, esse modelo vale-se de várias fontes de dados para descobrir o padrão de enchente, utilizando dados do nível rio como uma forma de calibração do modelo. Isso faz com que ele seja uma ferramenta importante para auxiliar os órgãos responsáveis, como a Defesa Civil, para alertar os moradores próximos das margens e para tomar todas as medidas preventivas, evitando problemas maiores que podem afetar toda uma cidade.

Figura 9 – Gráfico de analise de enchente do dia 23/11/15



Fonte: gráfico elaborado pelo autor.

4.2.2 Entendimento dos dados

Essa etapa é responsável pela compreensão de dados e a matéria-prima para que a solução seja construída. Dessa forma, o objetivo é conhecer as características e limitações das bases de dados, seu histórico, sua composição e seu tipo.

Um ponto relevante dessa parte é o entendimento da correlação entre os dados propriamente ditos. Sendo assim, foram escolhidos dados históricos que estão no mesmo período temporal de três anos - 2014, 2015 e 2016 -, exceto o mês de novembro de 2015, pois será usado para validação. Assim, esse modelo é desenvolvido a partir de uma mineração de dados que envolve dados do nível do rio do ponto 6 (ponto do Cristo) do projeto e-noe, que indicam quando ocorreram enchentes, juntamente com os dados da estações meteorológicas.

Os dados meteorológicos são importantes por serem fonte contínua de alimentação para o modelo depois de desenvolvido. Já o dado de nível do rio é utilizado só para a calibração do modelo, indicando quando se teve enchente ou não, a partir do momento que o modelo foi desenvolvido que e que captou o padrão dos dados meteorológicos antes, durante e depois da enchente. O modelo necessariamente precisará dos dados meteorológicos apenas para identificar enchentes. As variáveis das estações meteorológicas adquiridas para a região de São Carlos-SP e consideradas para o desenvolvimento desse modelo foram indicadas por meteorologista , sendo elas: *(i)* Temperatura Máxima, *(ii)* Temperatura Mínima, *(iii)* Umidade, *(iv)* Precipitação e *(v)* Intensidade do vento.

4.2.3 Preparação dos dados

Nessa etapa, são envolvidas todas as atividades associadas à construção do conjunto final de dados, aquele que será usado no treinamento do modelo, sofrendo inevitavelmente várias otimizações. Os dados foram modificados conforme o sugerido por (HAN; PEI; KAMBER, 2011). Todos esses passos foram possíveis utilizando a biblioteca *Pandas* [1] para a manipulação dos dados. Cada etapa foi importante em todas as fontes de dados, tanto na base de dados que mede o nível do rio quanto na base de dados das estações meteorológicas. A preparação dos dados passou pelos seguintes procedimentos:

- **Limpeza**: realiza a eliminação de caracteres espúrios, padronização de formatos, redução de inconsistência e imputação de dados faltantes. A frequência das bases de dados estavam diferentes. Enquanto os dados do nível do rio estavam com uma frequência de cinco em cinco minutos, os meteorológicos seguiam de frequência horária. Para que as características de inundação representadas nos dados do nível não fossem perdidas, os dados meteorológicos foram replicados em uma frequência de cinco minutos para alinharem-se àquelas do nível do rio, como descrito em (TSENG; WANG; LEE, 2003). Assim, as bases de dados foram padronizadas e estruturadas para que fosse possível entender os fenômenos anteriores, concomitantes e posteriores ao evento da enchente.

- **Transformação**: efetua a conversão de tipo, remoção de ruídos e o agrupamento de variáveis temporais, visando os períodos de antes, durante e depois das enchentes. Objetiva permitir que o modelo adquira o padrão desses períodos. Também houve normalização das variáveis e criação de novas variáveis como a *Label*, que indica se há enchente ou não, e a variável de tempo de concentração da bacia (TC), que estima empiricamente quanto tempo o ponto suporta o aumento de nível de água antes de inundar. Por fim, também demonstra a quantidade de chuva acumulada da precipitação.

  A *Label* que indica enchente ou não enchente - foi criada a partir do nível do rio (dados do e-noe). Esses dados são numéricos e representam a altura do rio. Sendo assim, 200 centímetros de altura são o limiar para indicar se há enchente naquele ponto de estudo, segundo especialistas que ajudaram na instalação do projeto. Ou seja, quando o rio ultrapassa os 200 centímetros, é considerada a possibilidade de enchente, adicionado assim a *Label* 1. Se a altura for inferior a 200 centímetros, não se considera a possibilidade de enchente, com a *Label* 0.

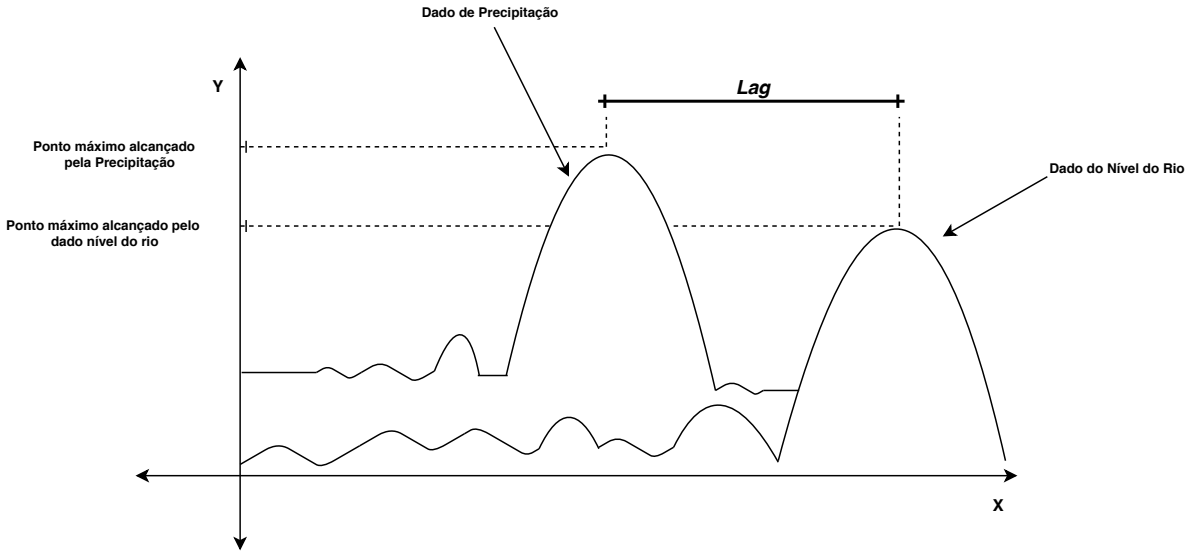
  Vale ressaltar que o ponto 6 do projeto e-noe (ponto do Cristo) foi o escolhido para estudo pois ocorrem muitas enchentes na área onde ele está instalado. Essa variável é essencial para o modelo na fase de treinamento, pois a técnica de classificação foi utilizada para

[1] Escrita para a linguagem de programação *Python* para a manipulação e análise de dados, Pandas é uma biblioteca *open source* amplamente utilizada na comunidade acadêmica. Em particular, oferece estruturas de dados e operações para manipular tabelas numéricas e séries temporais.

resolver o problema de enchentes. Essa técnica precisa de um exemplo tanto de enchente quanto de não enchente para, assim, criar um modelo, desenvolvendo a capacidade de classificar corretamente a partir de dados nunca vistos.

Como demostrado na Figura 10, a enchente não ocorre imediatamente quando começa a chover, mas é necessário um acúmulo de precipitação com o passar do tempo para que ela ocorra:

Figura 10 – Relação entre os dados do nível do rio e dados de precipitação



Fonte: figura elaborada pelo autor.

Para auxiliar o modelo a se adaptar nesse conceito, este trabalho a técnica de *Cross-Correlation (Correlação Cruzada)* [2] entre os dados de precipitação e os dados do nível do rio (ponto 6) no período de três anos de dados - 2014, 2015 e 2016 -, com o objetivo de encontrar o *Lag* entre elas, ou seja, o correlacionamento máximo entre as variáveis. O valor encontrado de *Lag* foi de 24, indicando assim a relação máxima existente entre elas durante esses três anos. Com esse valor, analisamos dois pontos:

– Conforme demostrado na Fórmula 4.1 onde **TC (Tempo de Concentração)** é a variável nova criada, **P** é o dado de precipitação, e **N** é o tamanho da base. A cada novo dado do TC (variável j indica dado diferente) é somado os últimos 24 valores dos dados históricos de precipitação (valor da *Lag*). Com isso o TC vai ser potencializado

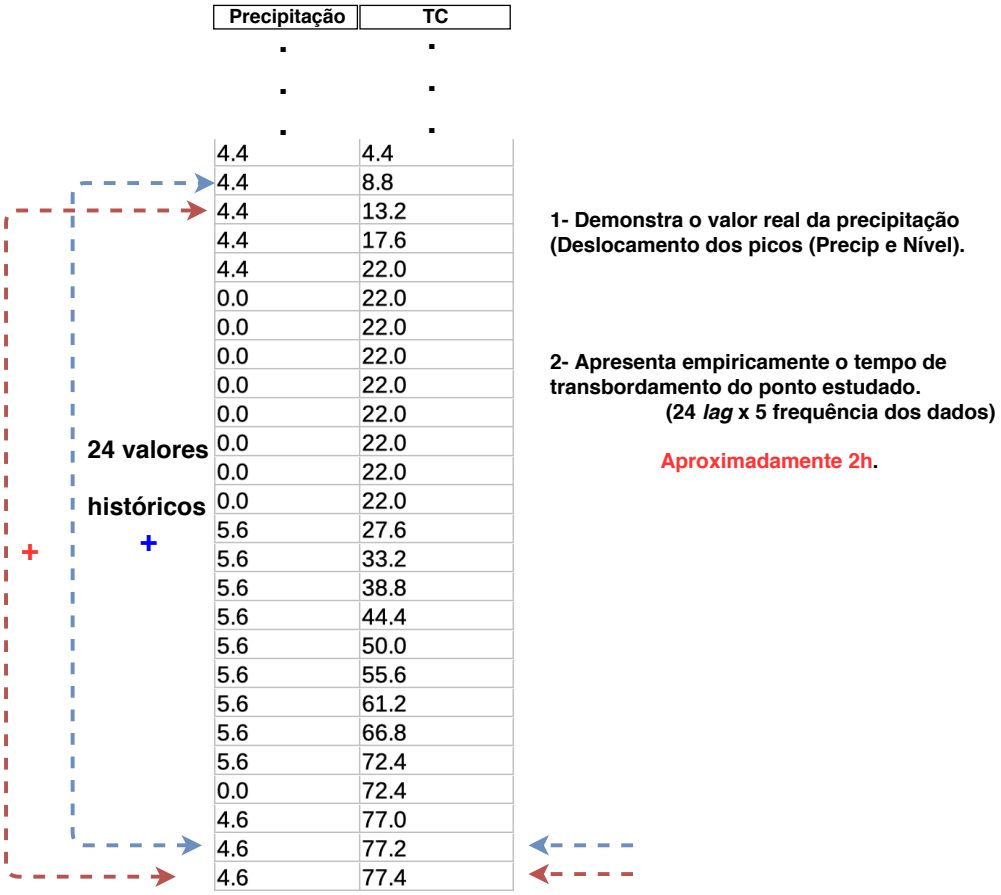
[2] A correlação cruzada é uma medida de similaridade de duas séries em função do deslocamento de uma em relação à outra. A função de correlação cruzada é útil para identificar defasagens da variável – atrasos (lags) ou avanços (leads) -, que podem ser preditores úteis da série principal.

para demostrar corretamente a quantidade de chuva que se acumulou a cada novo dado no tempo exato. Auxiliando, assim, a identificação de enchentes;

$$TC_j = \sum_{i=N-\mathbf{24}}^{N} P_i \tag{4.1}$$

– A Figura 11 demonstra como a variável TC é constituída numericamente e quais os ganhos que ela proporcionou ao modelo. A primeira característica é que ela apresenta de forma realista a quantidade exata de precipitação acumulada no ponto de estudo (ponto 6-ponto do Cristo). A variável TC também permitiu que fosse estimado empiricamente que, nesse ponto 6, a bacia leva aproximadamente duas horas até chegar ao ponto de transbordamento. Esse dado foi encontrado ao multiplicar por cinco o valor da *Lag* (24). Relembrando, cinco é a frequência modelada na base de dados final indicando os dados de precipitação e os outros dados de estações meteorológicas. Respeitando, então, a frequência disposta nos dados do nível do rio.

Figura 11 – Exemplo numérico variável TC e as características obtidas para o modelo

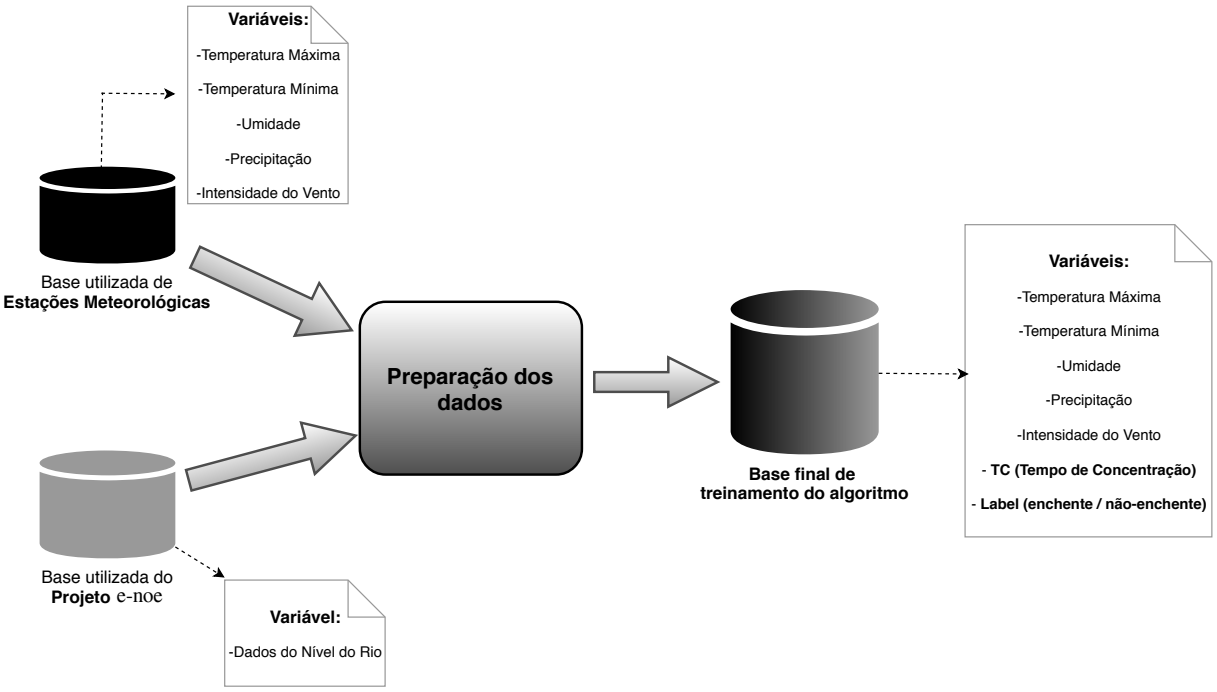


Fonte: figura elaborada pelo autor

Essa parte foi essencial para preparar os dados e, assim, melhorar a eficácia do modelo.

- **Integração das bases**: compilação das informações a partir de bases de dados distintas. Ao invés de usar de várias bases de dados, se constrói uma única com diversas variáveis. A Figura 12 mostra como a base era antes e como ficou após os devidos tratamentos necessários para o desenvolvimento do modelo.

Figura 12 – Base final para treinamento do algoritmo



Fonte: elaborada pelo autor

4.2.4 Modelagem, Avaliação e Aplicação do Modelo

A **modelagem** é responsável por identificar quais técnicas serão utilizadas para almejar o objetivo do projeto, seja ela uma predição, classificação, agrupamento ou regressão.

Nesse caso, a técnica de classificação tem como objetivo propor uma nova abordagem para a resolução desse problema em relação aos trabalhos do estado da arte, que o tratam com a técnica de regressão. Já a **avaliação** consiste em realizar testes e analisar os algoritmos para determinar quais deles obtiveram uma melhor performance e qual deles será utilizado no desenvolvimento do modelo. A seção de resultado traz, com maiores detalhes, quais avaliações foram feitas, quais algoritmos foram utilizados e qual algoritmo obteve melhor resultado.

A **aplicação do modelo** é o projeto já preparado e encerrado.

4.3 Resultados

Para a avaliação do modelo, a performance de cada algoritmo foi levada em consideração. Dessa forma, foram feitos testes tanto na ferramenta *Weka* como na biblioteca *scikit-learn* [3]. Como ambas obtiveram os mesmos resultados, a ferramenta *Weka* foi escolhida por obter uma maior validação na literatura e, assim, nela realizou-se a análise de cada algoritmo para posteriormente compará-los. Dentre todos os algoritmos disponíveis na ferramenta, foram avaliados cinco no âmbito de classificação.

As escolhas dos algoritmos foram baseadas em seus paradigmas e os parâmetros foram encontrados utilizando a biblioteca do Weka *CVParameterSelection* [4]. As informações dos algoritmos e os parâmetros utilizados são:

- Algoritmo baseado em probabilidade:
  - ***Naive Bayes***: não houve ajuste de parâmetro, pois ele é não paramétrico;
- Algoritmo baseado em função:
  - ***Multilayer Perceptron***: foi utilizada a taxa de aprendizado de 0.1, o momento foi de $0,8$ e o número de neurônios escondidos foi de 10;
  - ***SVM***: foi utilizado com a função *kernel* de base radial(RBF, do inglês *Radial Basis Function*). O *Gamma* foi de $0,75$ e a constante C de 1000;
- Algoritmo baseado em busca:
  - ***K-Nearest Neighbours (KNN):*** Os parâmetros utilizados foram o *default* do Weka versão 3-8-3;
- Algoritmo baseado em árvore de decisão:
  - ***Random_Forest:*** Foi utilizado *maxDepth* que indica a profundidade máxima da árvore com o valor de 10 e o *numIteration*, que indica a quantidade de árvores para interação, com o valor de 500. Para o restante dos parâmetros foi utilizado o *default* do Weka versão 3-8-3;

Dessa forma, o desempenho dos algoritmos foi analisado por meio do *k-fold cross-validation* com $k = 10$, sendo $k - 1$ para treino e o restante para teste, que é o padrão da ferramenta Weka. Assim, é possível gerar uma estimativa de erro mais precisa, pois a média das estimativas

[3] A scikit-learn é uma biblioteca de aprendizado de máquina de código aberto para a linguagem de programação Python.

[4] Biblioteca para otimizar a busca de melhores parâmetros por validação cruzada para qualquer classificador. Nela é dispostas uma lista de parâmetros e assim dentre eles é escolhido o com melhor performance.

tende a uma taxa de erro verdadeiro conforme o aumento de n e, geralmente, é utilizado para pequenos conjuntos de exemplos. Os resultados demonstram que o algoritmo *Random_Forest* possibilita uma classificação mais precisa se comparado aos outros classificadores selecionados. A Figura 13 apresenta o *Boxplots* referente às classificações realizadas. É importante ressaltar que o *Boxplot* com cor destacada refere-se aos resultados do *Randon_Forest* e exibe a mediana da acurácia superior alcançada pelos demais classificadores (os valores das medianas podem ser encontrados na Tabela 2 e a maior delas encontra-se destacada).

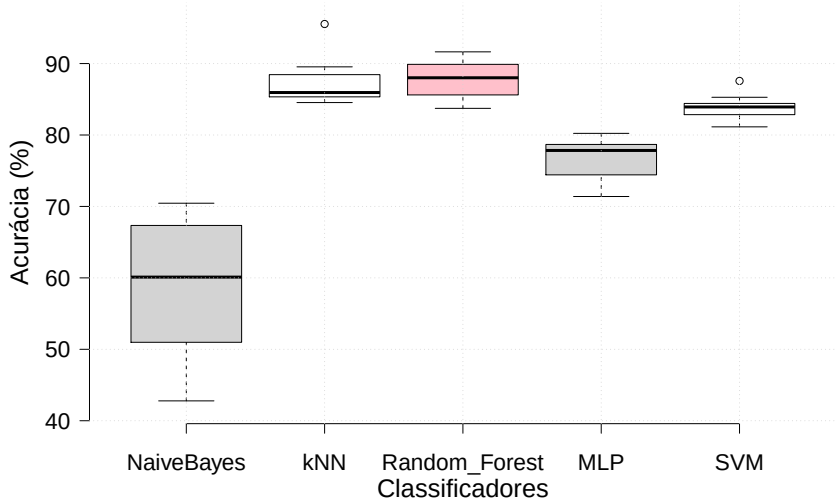


Figura 13 – Boxplots das acurácias apresentadas pelos classificadores para identificar enchentes. Tais resultados foram obtidos com o uso da técnica *k-fold cross-validation* com $k = 10$.

Com o objetivo de validar tais resultados, três análises estatísticas foram realizadas. Inicialmente, utilizou-se o método de *Shapiro Wilk* para verificar a sua adequação à normalidade e, consequentemente, para conduzir testes paramétricos ou não-paramétricos. *Shapiro Wilk* é um teste geral projetado para detectar todos os desvios de normalidade. O teste rejeita a hipótese de normalidade quando o valor p é menor ou igual a 0,05 (ROYSTON, 1992). Sendo assim, visto que os valores-p obtidos não foram todos maior que 0,05 (Tabela 2), é considerada recusada a hipótese de normalidade com confiabilidade de 95%. Portanto, o teste não-paramétrico é o mais indicado para as próximas análises.

Tabela 2 – Média (%) das acurácias e os valores-p dos conjuntos de resultados.

Classificadores	*Acurácia* Média (%)	*Shapiro Wilk* Valor-p
Naive Bayes	58,9	0.2966
kNN	87,1	**0.0031**
Random_Forest	**87,8**	0.9717
MLP	76,6	0.1755
SVM	83,9	0.7846

As comparações de pares realizadas com o teste *Wilcoxon Rank Sum* são exibidas na Tabela 3. O teste de *Wilcoxon* é um teste não paramétrico que compara dois grupos empare-

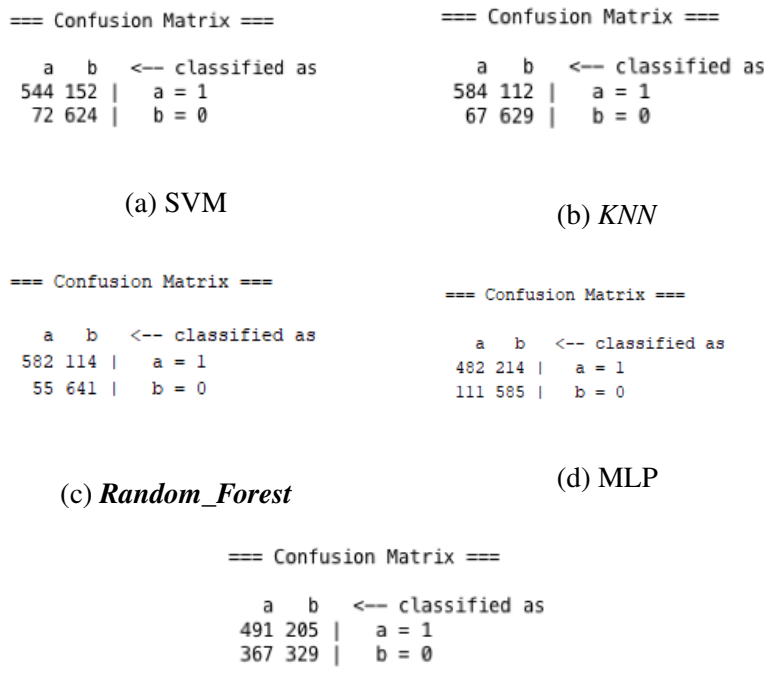
lhados. O teste essencialmente calcula a diferença entre cada conjunto de pares e analisa essas diferenças (WILCOXON, 1992). Os valores-*p* obtidos com a técnica de *Wilcoxon* indicam que somente os classificadores *Random_Forest* e *kNN* não apresentam diferença estatisticamente significativa entre si na classificação.

Tabela 3 – Valores-*p* da comparação de pares realizada com o teste *Wilcoxon Rank Sum*. Valores inferiores a 0,05 indicam diferença estatisticamente significativa entre os grupos de resultados.

	NaiveBayes	Random_Forest	kNN	MLP
Random_Forest	0.0001	-	-	-
kNN	0.0013	0.2411	-	-
MLP	0.0001	0.0001	0.0013	-
SVM	0.0013	0.0051	0.0051	0.0013

A segunda análise, foi feita a partir da utilização da métrica matriz de confusão, um tipo de tabela que permite a visualização do desempenho de um algoritmo de aprendizado. Ela permite visualizar a quantidade de erro e acerto de cada classe - Enchente (1) ou Não-enchente (0) - que o algoritmo obtém, de forma paralela (MONARD; BARANAUSKAS, 2003).

Figura 14 – *Confusion Matrix* (Matriz de confusão) dos algoritmos avaliados



(a) SVM

(b) *KNN*

(c) ***Random_Forest***

(d) MLP

(e) Naive Bayes

Fonte: elaborada pelo autor

Dessa forma, conforme a Figura 14, o algoritmo que se destacou e alcançou um maior acerto e um menor erro foi o ***Random_Forest***.

Para determinar qual o melhor algoritmo, uma análise minuciosa foi realizada a partir da medida precisão (*Precision*), cobertura (*Recall*) e *F-Measure*, oriundas da matriz de confusão. Essas métricas são definidas como:

- ***Precision**:* valor determinado estaticamente. Trata-se da precisão do modelo para a classificação correta das classes.
- ***Recall**:* o *Recall* considera a quantidade de informações daqueles dados classificados anteriormente pela *Precision* e leva em consideração a relevância desses classificados em relação a performance do modelo.
- ***F-Measure**:* Média harmônica entre *Precision* e o *Recall*. O resultado do *F-Measure* é um indicativo de que, quanto mais próximo de 1 melhor é o algoritmo. Do mesmo modo, quanto mais aproximados de 0, piores são os algoritmos.

A partir das análises, o algoritmo *Random_Forest* apresenta a maior acurácia média, com mais possibilidade de acerto e menos possibilidade de erro, além de possuir um maior *F-Measure*, como indicado na tabela 4. Esses resultados atestam que esse é o melhor algoritmo para o modelo, pois o que mediou a escolha do algoritmo para o desenvolvimento do modelo foi encontrar aquele que apresenta baixa taxa de erro nas análises. Isso significa que apresentará uma melhor performance em exemplos nunca vistos, o que pode mitigar os efeitos causados pelas enchentes e, futuramente, salvar vidas.

Tabela 4 – Tabela Comparativa entre os algoritmos

Algoritmo	Precision	Recall	F-Measure
MLP	0.772 ± 0.045	0.767 ± 0.083	0.765 ± 0.019
KNN	0.873 ± 0.027	0.871 ± 0.036	0.867 ± 0.034
SVM	0.843 ± 0.044	0.839 ± 0.065	0.838 ± 0.011
Random_Forest	0.881 ± 0.037	0.878 ± 0.048	0.878 ± 0.006
NaiveBayes	0.594 ± 0.0248	0.589 ± 0.13	0.583 ± 0.054

4.3.1 Simulação da Resiliência do Modelo

Esta subseção apresenta a potencialidade do modelo desenvolvido em relação à sua resiliência. Para determinar a capacidade de adaptação do modelo, alguns experimentos foram realizados. No primeiro deles, uma avaliação estatística foi feita mediante uma função intitulada *feature_importances* no *Random_Forest*. Essa função demonstra a importância de cada variável no tocante ao desenvolvimento do modelo – dividindo os dados em subconjuntos que mais pertencem a uma classe e, sem seguida, calculando matematicamente essas divisões para saber qual variável ajuda de forma mais eficaz a distinguir as classes. Assim, a saída dessa função é a importância de cada variável (em porcentagem) no contexto do modelo. Os resultados

obtidos foram: *(i)* Temperatura Máxima = 21%, *(ii)* Temperatura Mínima = 16%, *(iii)* Umidade = 24%, *(iv)* Precipitação= 10%, *(v)* Intensidade do vento = 4% e *(vi)* Tempo de Concentração (TC) = 22%.

Já no segundo experimento, uma simulação na ferramenta Weka foi realizada para que a performance e o comportamento do modelo desenvolvido fosse testados, levando em consideração a falta de variáveis em vista das adversidades que podem ocorrer no mundo real. A simulação é realizada seguindo os processos explicados mais pormenorizadamente nos próximos tópicos:

- **Modelo da simulação**: o modelo desenvolvido segue todos os passos do CRISP-DM, descritos nesta dissertação, e o algoritmo escolhido foi o *Random_Forest* por ter se destacado nos testes. Dessa forma, com o modelo treinado e validado, ele apenas necessitará dos dados meteorológicos (estações meteorológicas) e TC (Tempo de concentração) para identificar enchentes.
- **Dados da simulação**: os dados utilizados na simulação do modelo foram aqueles colhidos no mês de novembro de 2015. É importante ressaltar que esses dados não foram usados no treinamento do modelo.
- **Cenários**: com o objetivo de representar problemas que podem ocorrer no mundo real – como falhas de algumas variáveis essenciais das quais o modelo necessita para identificar enchentes –, surgiu a ideia de elaborar uma simulação a partir de cenários que representam possíveis perdas gradativas de variáveis, com a finalidade de verificar a influência existente na acurácia e no comportamento do modelo. Importante ressaltar que foram consideradas todas as combinações possíveis de variáveis para cada cenário. Sendo assim, foram desenvolvidos cinco cenários:
  - **Cenário** 1 (**C**1): apenas uma variável está disponível para o modelo para a identificação de enchentes;
  - **Cenário** 2 (**C**2): apenas duas variáveis estão disponíveis para o modelo, enquanto as outras estão indisponíveis;
  - **Cenário** 3 (**C**3): apenas três variáveis disponíveis para o modelo
  - **Cenário** 4 (**C**4): quatro variáveis disponíveis para o modelo, sendo que as outras estão indisponíveis;
  - **Cenário** 5 (**C**5): cinco variáveis disponíveis para o modelo, então apenas uma se encontra indisponível para identificar enchentes.

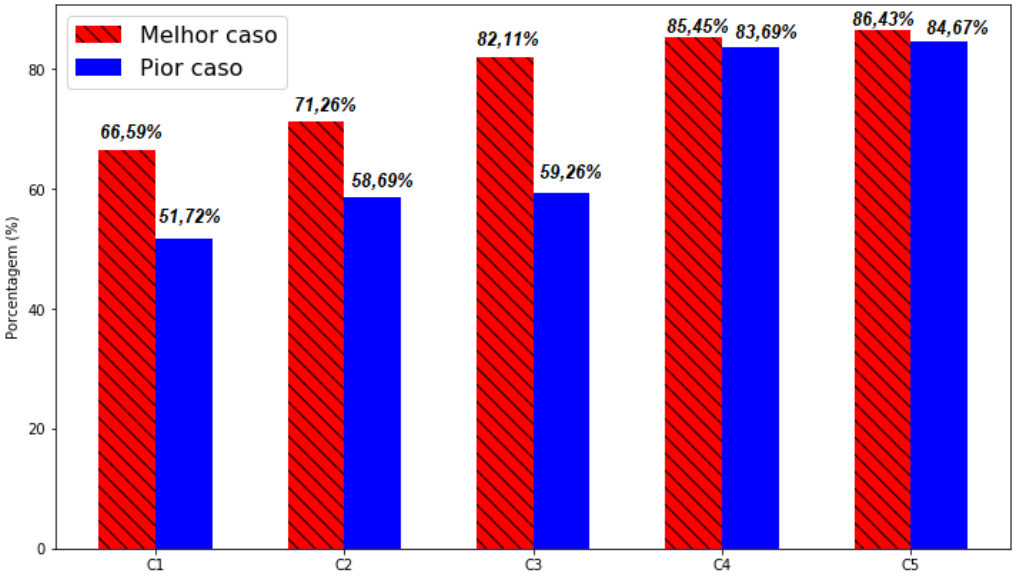
De maneira geral, a Figura 15 apresenta o comportamento obtido pelo modelo, baseada em sua acurácia mediante os cenários descritos. As variáveis que se destacaram na simulação para cada cenário são:

- **Cenário** 1 **(C**1**)**
  - ∗ **Melhor Caso:** variável de Umidade;
  - ∗ **Pior Caso:** variável de Intensidade do Vento;
- **Cenário** 2 **(C**2**)**
  - ∗ **Melhor Caso:** variáveis de Umidade e Tempo de Concentração (TC);
  - ∗ **Pior Caso:** variáveis de Intensidade do vento e Temperatura Mínima;
- **Cenário** 3 **(C**3**)**
  - ∗ **Melhor Caso:** variáveis de Temperatura Máxima, Umidade e Tempo de Concentração (TC);
  - ∗ **Pior Caso:** variáveis de Temperatura Mínima, Precipitação e Intensidade do Vento;
- **Cenário** 4 **(C**4**)**
  - ∗ **Melhor Caso:** variáveis de Temperatura Máxima, Umidade, Tempo de Concentração (TC) e Precipitação;
  - ∗ **Pior Caso:** variáveis de Intensidade do Vento, Temperatura Mínima, Tempo de Concentração (TC) e Precipitação;
- **Cenário** 5 **(C**5**)**
  - ∗ **Melhor Caso:** variáveis de Umidade, Temperatura Máxima, Tempo de Concentração (TC), Precipitação e Temperatura Mínima;
  - ∗ **Pior Caso:** variáveis de Intensidade do Vento, Tempo de Concentração (TC), Umidade e Temperatura Máxima e Precipitação;

Assim, podemos observar que o modelo se comportou bem na simulação, pois a perda significativa na acurácia está atrelada aos cenários 1 e 2, em que a média de diferença é de aproximadamente $18,87\%$ para o melhor caso e $32,59\%$ para o pior caso se comparada aos $87,8\%$ obtidos com todas as variáveis disponível. Entretanto, o funcionamento não é afetado, continuando a identificar enchentes. Sendo assim, trata-se de um ponto importante se comparado a outros trabalhos que não utilizam multivariáveis.

No caso dos cenários 3, 4 e 5, o modelo se comporta muito bem, pois a média de diferença de acurácia para o melhor caso é de $3,13\%$ e para o pior caso é de $11,92\%$ se comparado aos $87,8\%$ obtidos com todas as variáveis. Quando os dois experimentos são analisados em conjunto, fica claro que a variável de intensidade do vento é a que tem a menor porcentagem no quesito de importância da base de dados. Sendo assim, ela é a que mais é representada no pior caso, principalmente cenário 1. Já variáveis como temperatura mínima e precipitação se

Figura 15 – Resultados dos Cenários Propostos



Fonte: elaborada pelo autor.

encontram no centro entre o melhor caso e pior caso, em que necessitam de certas combinações para não influenciarem negativamente na acurácia do modelo. Já nos melhores casos dos cenários, variáveis como umidade, tempo de concentração e temperatura máxima aparecem como valores que se combinam para auxiliar a eficácia do modelo no contexto dos cenários propostos. Portanto, percebe-se que, mesmo a partir das adversidades de todos os cenários o modelo não deixou de funcionar e, mesmo que haja a indisponibilidade de metade das variáveis relatada no cenário 3, obtém-se ainda uma acurácia de $82,45\%$ no melhor caso, ficando muito próxima da acurácia de $87,8\%$ obtida com todas as variáveis. Esses dois fatores demonstram a capacidade de adaptabilidade do modelo mediante as adversidades. Essa característica é bastante importante quando o assunto são os desastres naturais, tornando esse modelo uma ferramenta essencial para identificação de enchentes.

4.3.2 Comparação do Modelo Proposto X Storm Water Management Model (SWMM)

Na literatura, modelos hidrológicos vêm sendo amplamente utilizados para a previsão e alerta de desastres por enchentes. Existem diversos modelos com diferentes configurações e requerimento de dados para seu funcionamento. Como descrito na fundamentação teórica, cada categoria de modelo de previsão de enchentes possui recursos que devem ser levados em consideração para a escolha do modelo a ser adotado como, por exemplo, a simplicidade de uso, eficácia dos resultados, custos computacionais e complexidade. Assim, foi utilizado para comparação o modelo *Storm Water Management Model* (SWMM), descrito por (FAVA *et al.*,

2018), que realiza o monitoramento e a previsão de enchentes na bacia urbana de São Carlos-SP. A escolha dele se dá por ser um modelo introduzido na mesma região de estudos deste trabalho.

O modelo SWMM é bastante explorado na literatura para diversos tipos de estudo, tanto quantitativos como qualitativos. Ele é um modelo hidrodinâmico semi-distribuído que possui componentes hidráulicos e hidrológicos. O trabalho realizado para a área de São Carlos utilizou os dados de nível do projeto e-noe e uma rede experimental de quatro pluviógrafos como entrada para as simulações realizadas. No entanto, além dos dados de nível e chuva, esse modelo também requer um grande detalhamento do uso e ocupação do solo, características hidráulicas da bacia, entre outros dados. Esses parâmetros precisam ser calibrados e validados para que se tenha uma performance de previsão e cenários de chuva extremamente confiáveis. Apesar do modelo ser bastante robusto e ser capaz de descrever o comportamento da bacia de forma distribuída durante um evento de chuva, existe também a desvantagem de requerer um enorme detalhamento da bacia, além de grande esforço para a calibração e ajuste de todos os seus parâmetros. Outro ponto crítico em relação ao modelo distribuído, se comparado a um modelo empírico, é o custo computacional. O modelo semi-distribuído considera muitas características físicas da bacia, peculiaridades da área, drenagem e uso do solo, o que acaba tornando-o custoso computacionalmente e, por sua vez, dificultando a simulação em tempo real para locais em que a onda de cheia acontece repentinamente.

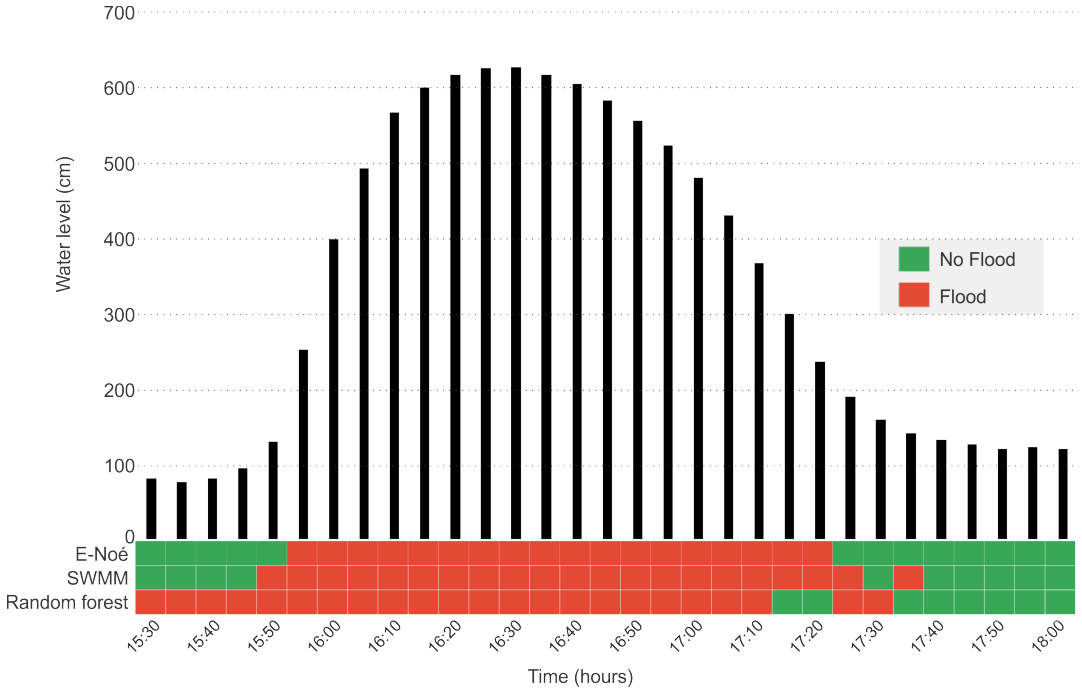


Figura 16 – Gráfico comparação do dia 23/11/15.

Considerando este modelo semi-distribuído, aplicado na mesma área de estudo deste trabalho, é importante avaliar e comparar seus resultados com o método desenvolvido. O modelo SWMM gera saídas quantitativas. Para fins de comparação com o modelo desenvolvido, as saídas produzidas pelo SWMM foram convertidas em respostas binárias. A figura 16 mostra

os resultados de identificação de enchente pelo modelo SWMM e pelo método proposto para o evento de chuva ocorrido no dia 23 de novembro de 2015, comparados com as respostas medidas pelo sensor. Pela figura, podemos observar que ambos os modelos foram capazes de identificar a condição de enchente para esse evento. Ambos os modelos superestimaram o evento e detectaram a enchente antes mesmo de ela estar realmente acontecendo. No entanto, o erro não passa de 25 minutos de adiantamento pelo método desenvolvido e cinco minutos pelo modelo SWMM. Esses resultados demonstram que o modelo proposto identificou com 25 minutos de antecedência a enchente. Essa informação poderia ser utilizada para gerar alertas aos órgãos responsáveis e assim mitigar os efeitos causados pelas enchentes. Além da comparação visual desse evento específico, uma outra comparação também foi realizada, agora para perceber a eficácia para toda a série de dados (mês de novembro). A tabela 5 mostra que os dois modelos a apresentaram resultados satisfatórios com eficácia de $99,49\%$ e $99,6\%$, valores do modelo desenvolvido e do SWMM, respectivamente. Para a série de 8259 dados, o modelo SWMM apresentou apenas 33 erros de falso positivo, ou seja, detectou enchentes quando não estavam ocorrendo. Já o método desenvolvido teve 42 erros, acusando cinco falsos positivos e 37 falsos negativos. Segundo os hidrólogos, os erros relacionados aos falsos negativos representam a credibilidade do modelo, pois para eles é melhor um modelo que tenha mais erros falsos negativos do que erros falsos positivos, pois isso garante que o modelo identificou com antecedência o evento da enchente.

Tabela 5 – Tabela Comparativa da performance dos modelos

Modelos	**Modelo Desenvolvido (Random forest)**	**SWMM**
Nº de Dados	8259	8259
Acertos(%)	8217 (99,50%)	8226 (99,60%)
Erros	42 (0,50%)	33 (0,4%)
Falsos (+)	5 (11,91% de 42 erros)	33 (100% dos 33 erros)
Falsos (-)	37 (88,09% de 42 erros)	0

De forma geral, o desempenho do modelo SWMM foi ligeiramente mais positivo. Entretanto, considerando a complexidade e o custo para o seu desenvolvimento – e principalmente tendo uma resolução mais alta para representar os recursos, requerendo maior quantidade de informações –, acaba induzindo um tempo computacional maior em comparação ao modelo proposto. O modelo desenvolvido obteve resultados extremamente satisfatórios, mesmo sendo considerado um modelo empírico. Segundo o trabalho de (DUNCAN *et al.*, 2011), o tempo computacional de um modelo empírico é mais curto em relação aos com modelos convencionais baseados em física, como é o caso do SWMM, pois a sua resolução é mais baixa o que resulta em uma menor quantidade de dados para seu desenvolvimento. Consequentemente, esse modelo torna-se menos complexo e custoso. Portanto, como demostrado na revisão da literatura, inúmeras cidades no Brasil se encontram vulneráveis e susceptíveis à ocorrência de enchentes e contam com pouco ou nenhum dado disponível para o monitoramento de enchentes. Logo, a instalação de um modelo empírico, como o proposto neste trabalho, é bastante vantajosa, pois os dados requeridos são apenas aqueles medidos através de um sensor de nível e dados de chuva, fornecidos

pelas estações meteorológicas. Esses dados de chuva estão disponíveis para todo o território nacional. Assim, o método tem alta replicabilidade para qualquer região do país, pois não requer a instalação de uma extensa rede de monitoramento e obtém uma performance aceitável no contexto de identificação de enchentes.

CAPÍTULO

5

CONCLUSÃO

As enchentes são eventos que ocorrem em todo o mundo. Devido às mudanças climáticas, sabemos que não são possíveis de serem evitadas. Entretanto, com o uso da computação e tecnologia, é possível gerenciar e possibilitar às cidades que se preparem durante os eventos de chuvas. Este trabalho demonstra o desenvolvimento de um modelo baseado em mineração de dados em diversas bases com aprendizado de máquina, com a finalidade de tornar possível a identificação de enchentes e, assim, mitigar o máximo possível os efeitos que elas podem causar. Dessa forma, essa abordagem entra no conceito de cidades inteligentes, ou seja, cidades monitoradas em todos os seus aspectos com o objetivo de melhorar sua infraestrutura. Essa monitoração é realizada em tempo real utilizando diversas tecnologias e melhorando a qualidade de vida em centros urbanos. Os trabalhos das literaturas da área, principalmente os modelos hidrológicos, são habitualmente utilizados para a identificação de enchentes. Entretanto, a maioria dos modelos é bastante complexa e custosa em longo prazo. Isso se deve ao fato de necessitarem de grande quantidade de dados da bacia hidrográfica da região, tornando a calibração uma tarefa não-trivial. Outro ponto é a recalibração necessária, pois as variáveis sofrem mudanças com o passar do tempo, resultando na perda de eficiência desses modelos.

Esta dissertação explora o desenvolvimento de um modelo de identificação de enchentes que utiliza mineração de dados em várias bases de dados (dados do nível do rio correlacionados aos dados de estações meteorológicas) para, de fato, encontrar o padrão das enchentes a partir do aprendizado de máquina. O trabalho tem o importante papel de fortalecer cada vez mais a ideia de um ambiente inteligente que utiliza aprendizado de máquina e diversas bases de dados para tomar decisões.

Resultados mostram que o algoritmo *Random_Forest* foi o que obteve maior aceitação na avaliação realizada juntamente com outros algoritmos, alcançando uma taxa de acurácia de aproximadamente 87.8% e passando com probidade em relação às análises realizadas. Esse resultado demonstra que o modelo desenvolvido cumpre com a premissa de que o correlaciona-

mento de mineração de dados, as diversas bases de dados e o aprendizado de máquina podem ser ferramentas importantes no contexto de cidades inteligentes.

5.1 Síntese das Contribuições

Em vista das principais contribuições desta dissertação, podemos ressaltar:

- Este trabalho demostra um avanço no estado da arte, pois desenvolveu um modelo que identifica enchentes sem necessitar do dado do nível do rio, visto que (FURQUIM *et al.*, 2018) existe uma dependência integral da variável do nível do rio para essa identificação. O modelo desenvolvido tem uma dependência parcial e utiliza o dado do rio como *label* para calibração na fase de treinamento. Após o treinamento, o modelo não necessita mais do dado do nível do rio, pois identifica enchente a partir das outras variáveis dispostas na base, constituindo-se como um modelo próximo do tempo real;

- Um outro ponto é a complexidade e o custo para o desenvolvimento dos modelos hidrológicos, considerando a grande quantidade de dados necessária para a calibragem e a recalibragem em longo prazo. Um exemplo é o trabalho de (FAVA, 2015), que desenvolveu um modelo hidrológico para a mesma região deste estudo. O tempo do desenvolvimento do modelo, por ser muito extenso, também acaba sendo comprometido ao longo de todo esse processo. O modelo desenvolvido diminui a complexidade e o custo com técnicas de aprendizado de máquinas para encontrar um padrão nos dados e calibrá-los com as variáveis que são adquiridas mais facilmente. Além disso, o modelo necessita de uma única calibragem no treinamento – o que possibilita diminuir o processo de desenvolvimento em relação aos modelos hidrológicos, além deste apresentar uma mais resiliência em longo prazo.

- Outra contribuição notável do modelo proposto é em relação à variável de tempo de concentração(TC) da bacia nos modelos hidrológicos, pois é encontrada empiricamente a partir de fórmulas matemáticas. O trabalho (FAVA *et al.*, 2018) cita o valor encontrado para o ponto 6 (ponto do Cristo) a partir do método de *Kirpish*, maneira não-trivial e que torna necessários vários cálculos para encontrá-lo empiricamente. Já no modelo desta dissertação, o TC é encontrado computacionalmente a partir do *Cross Correlation* entre o dado do rio e a dado de precipitação, em forma de *lag*. Assim, o valor encontrado empiricamente é próximo àquele do método de *Kirpish* mas de forma menos custosa e mais facilmente generalizável para uma região.

5.2 Limitações e Trabalhos Futuros

Após a realização deste trabalho, foram encontradas algumas limitações e possibilidades de trabalhos futuros, a saber:

- No tocante às limitações, a obtenção dos dados nas estações meteorológicas deveriam ser mais adaptativas para o monitoramento do ambiente, com a mudança de frequência baseada no evento do desastre. Isso auxiliaria a construção do modelo, facilitando, assim, a padronização dos dados com outras bases e podendo representar um ganho de eficácia, pois utilizaria dados mais consistentes e realistas para o monitoramento.

- Outra limitação refere-se à quantidade de dados históricos, necessários para o treinamento do modelo. Uma solução para os trabalhos futuros seria a utilização de técnicas, como o aprendizado de máquina sem fim ou até um aprendizado online. Técnicas que permitem o aprendizado constante, sem necessitar de uma grande quantidade de dados, fazendo com que o modelo aprenda gradativamente à medida que os dados vão sendo obtidos.

- Outra sugestão para os trabalhos futuros é a utilização de outros tipos de sensores, como dados de satélite, dados de raios, entre outros, pois possibilitam um ganho de acurácia. Além disso, construir um modelo mais generalizável, considerando dados de outros pontos do projeto e-noe, ao invés de apenas um como descrito nesta dissertação.

- Por fim, considerar, em trabalhos futuros, uso desta metodologia para também identificar outros tipos de desastres.

5.3 Publicações e Trabalhos em Andamento

Artigo aceito: **BRITO, L. A. V.**; BRESSIANI, D. ; UEYAMA, J. Explorando aprendizado de máquina com multivariáveis para previsão de enchentes em ambientes IoTs : um estudo empírico no sistema de monitoramento de rios E-noé. In: II WORKSHOP DE COMPUTAÇÃO URBANA, 2018, Campos do Jordão. Anais do II Workshop de Computação Urbana, 2018. Porto Alegre: Sociedade Brasileira de Computação (SBC), 2018. v. 2, n. 1/2018, may 2018. ISSN 2595-2706

Artigo em andamento: **BRITO, L. A. V.**; ANDRADE, S. ;MANO, L.; FAVA, M.; MEDIONDO E; UEYAMA, J. Modelo de Classificação Multivariável para Identificação de Enchentes: um estudo empírico no sistema de monitoramento de rios e-noe. **Journal**: *Journal of Hyfroinformatics* **Fator de Impacto**: 1.797

REFERÊNCIAS

ACOSTA-COLL, M.; BALLESTER-MERELO, F.; MARTÍNEZ-PEIRÓ, M. Early warning system for detection of urban pluvial flooding hazard levels in an ungauged basin. **Natural Hazards**, Springer, v. 92, n. 2, p. 1237–1265, 2018. Citado nas páginas 41 e 43.

AKYILDIZ, I. F.; SU, W.; SANKARASUBRAMANIAM, Y.; CAYIRCI, E. A survey on sensor networks. **IEEE Communications magazine**, IEEE, v. 40, n. 8, p. 102–114, 2002. Citado nas páginas 35 e 36.

ALMEIDA, L.; SERRA, J. C. V. Modelos hidrológicos, tipos e aplicações mais utilizadas. **Revista da FAE**, v. 20, n. 1, p. 129–137, 2017. Citado nas páginas 37 e 38.

AMO, S. D. Técnicas de mineração de dados. **Jornada de Atualizaçao em Informatica**, 2004. Citado na página 29.

BOTH, G. C.; HAETINGER, C.; FERREIRA, E. R.; DIEDRICH, V. L.; AZAMBUJA, J. Fay de. Uso da modelagem matemática para a previsão de enchentes no vale do taquari–rs. **Anais do Simpósio Brasileiro de Engenharia Ambiental [CD-ROM]**, 2008. Citado nas páginas 42 e 43.

BRAGA, A. de P. **Redes neurais artificiais: teoria e aplicações**. LTC Editora, 2007. ISBN 9788521615644. Disponível em: <https://books.google.com.br/books?id=R-p1GwAACAAJ>. Citado na página 33.

BRITO, L. A. V.; BRESSIANI, D.; UEYAMA, J. Explorando aprendizado de máquina com multivariáveis para previsao de enchentes em ambientes iots: um estudo empırico no sistema de monitoramento de rios e-noé. **Anais do II Workshop de Computação Urbana (COURB 2018)**, v. 2, n. 1/2018, 2018. Citado na página 46.

BROWN, M. S. **Data mining for dummies**. [S.l.]: John Wiley & Sons, 2014. Citado nas páginas 27 e 45.

CARVALHO, F. B. S. de; LEAL, B. G.; FILHO, J. V. dos S.; BAIOCCHI, O. R.; LOPES, W. T.; ALENCAR, M. S. de. Aplicacoes ambientais de redes de sensores sem fio. **Revista de tecnologia da informação e comunicação**, v. 2, n. 1, p. 14–19, 2012. Citado na página 24.

CHEN, D.; LIU, Z.; WANG, L.; DOU, M.; CHEN, J.; LI, H. Natural disaster monitoring with wireless sensor networks: a case study of data-intensive applications upon low-cost scalable systems. **Mobile Networks and Applications**, Springer, v. 18, n. 5, p. 651–663, 2013. Citado nas páginas 42 e 43.

CREPALDI, P. G.; AVILA, R. N. P.; PAULO, J. P. M. de O.; RODRIGUES, R.; MARTINS, R. L. Um estudo sobre a árvore de decisão e sua importância na habilidade de aprendizado. **Revista Eletrônica do Instituto de Ensino Superior de Londrina**, 2011. Citado na página 35.

DEVIA, G. K.; GANASRI, B.; DWARAKISH, G. A review on hydrological models. **Aquatic Procedia**, Elsevier, v. 4, p. 1001–1007, 2015. Citado na página 38.

DUNCAN, A.; CHEN, A. S.; KEEDWELL, E.; DJORDJEVIC, S.; SAVIC, D. Urban flood prediction in real-time from weather radar and rainfall data using artificial neural networks. **International Association of Hydrological Sciences**, 2011. Citado na página 61.

EHLERS, R. S. Introduçao a inferência bayesiana. **Disponıvel em< http://www. icmc. usp. br/ehlers/notas/bayes. pdf>. Acesso em**, v. 1, n. 03, p. 2009, 2007. Citado na página 32.

FAVA, M. C. **Modelo de alerta hidrológico com base participativa usando sistema de informações voluntárias para previsão de enchentes**. Dissertação (Mestrado) — Escola de Engenharia de São Carlos, Universidade de São Paulo, São Carlos, São Carlos, 2015. Citado nas páginas 42, 43 e 64.

FAVA, M. C.; MAZZOLENI, M.; ABE, N.; MEDIOND, E. M.; SOLOMATINE, D. P. An approach for urban catchment model updating. 2018. Citado nas páginas 25, 60 e 64.

FREITAS, C. M. d.; SILVA, D. R. X.; SENA, A. R. M. d.; SILVA, E. L.; SALES, L. B. F.; CARVALHO, M. L. d.; MAZOTO, M. L.; BARCELLOS, C.; COSTA, A. M.; OLIVEIRA, M. L. C. *et al.* Desastres naturais e saúde: uma análise da situação do brasil. **Ciência & Saúde Coletiva**, SciELO Public Health, v. 19, p. 3645–3656, 2014. Citado na página 23.

FUKUNAGA, K.; NARENDRA, P. M. A branch and bound algorithm for computing k-nearest neighbors. **IEEE transactions on computers**, IEEE, v. 100, n. 7, p. 750–753, 1975. Citado na página 34.

FURQUIM, G.; JALALI, R.; PESSIN, G.; PAZZI, R. W.; UEYAMA, J. *et al.* How to improve fault tolerance in disaster predictions: a case study about flash floods using iot, ml and real data. **Sensors**, Multidisciplinary Digital Publishing Institute, v. 18, n. 3, p. 907, 2018. Citado nas páginas 41, 43 e 64.

FURQUIM, G. A. **Uma abordagem tolerante a falhas para a previsão de desastres naturais baseada em IoT e aprendizado de máquina**. Tese (Doutorado) — USP, São Carlos, 8 2017. Citado na página 47.

GARDNER, M. W.; DORLING, S. Artificial neural networks (the multilayer perceptron)—a review of applications in the atmospheric sciences. **Atmospheric environment**, Elsevier, v. 32, n. 14-15, p. 2627–2636, 1998. Citado na página 33.

____. Artificial neural networks (the multilayer perceptron)—a review of applications in the atmospheric sciences. **Atmospheric environment**, Elsevier, v. 32, n. 14-15, p. 2627–2636, 1998. Citado na página 33.

HALL, M.; FRANK, E.; HOLMES, G.; PFAHRINGER, B.; REUTEMANN, P.; WITTEN, I. H. The weka data mining software: an update. **ACM SIGKDD explorations newsletter**, ACM, v. 11, n. 1, p. 10–18, 2009. Citado na página 39.

HAN, J.; PEI, J.; KAMBER, M. **Data mining: concepts and techniques**. [S.l.]: Elsevier, 2011. Citado na página 49.

HO, T. K. Random decision forests. In: IEEE. **Document analysis and recognition, 1995., proceedings of the third international conference on**. [S.l.], 1995. v. 1, p. 278–282. Citado na página 34.

KEMP, K. K. Environmental modeling with gis: A strategy for dealing with spatial continuity. **National Center for Geographic Information and Analysis (NCGIA)**, Citeseer, 1993. Citado na página 24.

KOTSIANTIS, S. B.; ZAHARAKIS, I.; PINTELAS, P. Supervised machine learning: A review of classification techniques. **Emerging artificial intelligence applications in computer engineering**, v. 160, p. 3–24, 2007. Citado na página 39.

KURGAN, L. A.; MUSILEK, P. A survey of knowledge discovery and data mining process models. **The Knowledge Engineering Review**, Cambridge University Press, v. 21, n. 1, p. 1–24, 2006. Citado na página 45.

LEMOS, A. Cidades inteligentes. **GV-executivo**, v. 12, n. 2, p. 46–49, 2013. Citado na página 24.

LINOFF, G. S.; BERRY, M. J. **Data mining techniques: for marketing, sales, and customer relationship management**. [S.l.]: John Wiley & Sons, 2011. Citado na página 29.

LORENA, A. C.; CARVALHO, A. C. de. Uma introdução às support vector machines. **Revista de Informática Teórica e Aplicada**, v. 14, n. 2, p. 43–67, 2007. Citado na página 33.

LOUREIRO, A. A.; NOGUEIRA, J. M. S.; RUIZ, L. B.; MINI, R. A. d. F.; NAKAMURA, E. F.; FIGUEIREDO, C. M. S. Redes de sensores sem fio. In: **Simpósio Brasileiro de Redes de Computadores**. [S.l.: s.n.], 2003. v. 21, p. 19–23. Citado nas páginas 24 e 35.

MITCHELL, T. M. **Machine Learning**. 1. ed. New York, NY, USA: McGraw-Hill, Inc., 1997. ISBN 0070428077, 9780070428072. Citado na página 31.

MONARD, M. C.; BARANAUSKAS, J. A. Conceitos sobre aprendizado de máquina. **Sistemas Inteligentes-Fundamentos e Aplicações**, v. 1, n. 1, p. 32, 2003. Citado na página 55.

MOSTAFA, E.; MOHAMED, E. *et al.* Intelligent data classification and aggregation in wireless sensors for flood forecasting system. In: IEEE. **Microwave Symposium (MMS), 2014 14th Mediterranean**. [S.l.], 2014. p. 1–8. Citado nas páginas 41 e 43.

NASCIMENTO, R. F. F.; ALCÂNTARA, E.; KAMPEL, M.; STECH, J. L.; NOVO, E.; FONSECA, L. M. G. O algoritmo support vector machines (svm): avaliação da separação ótima de classes em imagens ccd-cbers-2. **Simpósio Brasileiro de Sensoriamento Remoto**, v. 14, p. 2079–2086, 2009. Citado na página 34.

ONU. Implementation of the international strategy for disaster reduction. In: GENERAL ASSEMBLY. **Sustainable development: International Strategy for Disaster Reduction**. 2014. Disponível em: <http://https://www.unisdr.org/files/resolutions/N1452549.pdf>. Acesso em: 20 agost. 2018. Citado na página 23.

PECHOTO, M. M.; UEYAMA, J.; PEREIRA, J. E-noé: Rede de sensores sem fio para monitorar rios urbanos. In: **Congresso Brasileiro Sobre Desastres Naturais**. [S.l.: s.n.], 2012. Citado nas páginas 24, 36 e 37.

REIS, L. P.; VIEIRA, J.; LEMOS, P.; NOVAIS, R.; FARIA, B. M. Higher education access prediction using data-mining. In: IEEE. **Information Systems and Technologies (CISTI), 2017 12th Iberian Conference on**. [S.l.], 2017. p. 1–8. Citado nas páginas 29, 31 e 46.

ROYSTON, P. Approximating the shapiro-wilk w-test for non-normality. **Statistics and Computing**, Springer, v. 2, n. 3, p. 117–119, 1992. Citado na página 54.

SANTOS, L. L. Modelos hidrológicos: Conceitos e aplicações. **Revista Brasileira de Geografia Física**, v. 2, n. 3, p. 1–19, 2009. Citado na página 24.

SILVA, R. T.; PORTO, M. F. d. A. Gestão urbana e gestão das águas: caminhos da integração. **Estudos avançados**, SciELO Brasil, v. 17, n. 47, p. 129–145, 2003. Citado na página 23.

SOUZA, A. S.; CURVELLO, A. M. de L.; SOUZA, F. L. d. S. de; SILVA, H. J. da. A flood warning system to critical region. **Procedia Computer Science**, Elsevier, v. 109, p. 1104–1109, 2017. Citado na página 23.

Souza, F. A. A. D.; Mendiondo, E. M.; Taffarello, D.; Guzmán-Arias, D.; Fava, M. C.; Abreu, F.; Freitas, C. C.; de Macedo, M. B.; Estrada, C. R.; do Lago, C. A. Socio-Hydrological Observatory for Water Security (SHOWS): Examples of Adaptation Strategies With Next Challenges from Brazilian Risk Areas. **AGU Fall Meeting Abstracts**, dez. 2018. Citado na página 23.

TSENG, S.-M.; WANG, K.-H.; LEE, C.-I. A pre-processing method to deal with missing values by integrating clustering and regression techniques. **Applied Artificial Intelligence**, Taylor & Francis, v. 17, n. 5-6, p. 535–544, 2003. Citado na página 49.

TUCCI, C. E.; HESPANHOL, I.; NETTO, O. d. M. C. Cenários da gestão da água no brasil: uma contribuição para a "visão mundial da água". **Interações**, v. 1980, p. 90, 2003. Citado nas páginas 23 e 37.

TUCCI, C. E. *et al.* Modelos hidrológicos. **UFRGS/Associação Brasileira de recursos Hídricos**, 1998. Citado nas páginas 24 e 25.

VAPNIK, V. **The nature of statistical learning theory**. [S.l.]: Springer science & business media, 2013. Citado na página 33.

VIEIRA, H. de F. **Provendo resiliência em uma rede de sensores sem fio linear e esparsa através de veículo aéreo não tripulado**. 58 f. Tese (Doutorado) — Intituto de Ciências Matemáticas e de Computação de São Carlos (ICMC/USP), Universidade de São Paulo, São Carlos - SP, 2015. Disponível em: <http://http://www.teses.usp.br/teses/disponiveis/55/55134/tde-14082015-103230/en.php>. Acesso em: 15 jul. 2017. Citado nas páginas 37 e 38.

WEISS, S. M.; INDURKHYA, N. **Predictive data mining: a practical guide**. [S.l.]: Morgan Kaufmann, 1998. Citado na página 29.

WILCOXON, F. Individual comparisons by ranking methods. In: **Breakthroughs in Statistics**. [S.l.]: Springer, 1992. p. 196–202. Citado na página 55.

YIN, J.; LAMPERT, A.; CAMERON, M.; ROBINSON, B.; POWER, R. Using social media to enhance emergency situation awareness. **IEEE Intelligent Systems**, v. 27, n. 6, p. 52–59, 2012. Citado na página 24.


ICMC
USP
SÃO CARLOS