UNIVERSIDADE TECNOLÓGICA FEDERAL DO PARANÁ
PROGRAMA DE PÓS-GRADUAÇÃO EM COMPUTAÇÃO APLICADA

WILLIAN MUNIZ DO NASCIMENTO

DETECÇÃO DE VAZAMENTOS EM DADOS DE FLUXO DE ÁGUA COM SELEÇÃO E OTIMIZAÇÃO AUTOMÁTICA DE MODELOS

DISSERTAÇÃO DE MESTRADO

CURITIBA
2021

WILLIAN MUNIZ DO NASCIMENTO

DETECÇÃO DE VAZAMENTOS EM DADOS DE FLUXO DE ÁGUA COM SELEÇÃO E OTIMIZAÇÃO AUTOMÁTICA DE MODELOS

Leakage Detection using Water Flow Data with Automatic Model Selection and Optimization

Dissertação de Mestrado apresentado(a) como requisito parcial à obtenção do título de Mestre em Computação Aplicada, do Programa de Pós-Graduação em Computação Aplicada, da Universidade Tecnológica Federal do Paraná.

Orientador: Prof. Dr. Luiz Celso Gomes Junior

CURITIBA

2021

Ministério da Educação
Universidade Tecnológica Federal do Paraná
Campus Curitiba

WILLIAN MUNIZ DO NASCIMENTO

DETECÇÃO DE VAZAMENTOS EM DADOS DE FLUXO DE ÁGUA COM SELEÇÃO E OTIMIZAÇÃO AUTOMÁTICA DE MODELOS

Trabalho de pesquisa de mestrado apresentado como requisito para obtenção do título de Mestre Em Computação Aplicada da Universidade Tecnológica Federal do Paraná (UTFPR). Área de concentração: Engenharia De Sistemas Computacionais.

Data de aprovação: 16 de Agosto de 2021

Prof Luiz Celso Gomes Junior, Doutorado - Universidade Tecnológica Federal do Paraná

Prof Marcelo Dalcul Depexe, Mestrado - Sanepar

Prof.a Myriam Regattieri De Biase Da Silva Delgado, Doutorado - Universidade Tecnológica Federal do Paraná

Prof Rodrigo Jardim Riella, Doutorado - Instituto de Tecnologia para Desenvolvimento, Departamento de Eletricidade, Divisão de Sistemas Elétricos - Lactec

Documento gerado pelo Sistema Acadêmico da UTFPR a partir dos dados da Ata de Defesa em 20/08/2021.

Dedico este trabalho a minha família,
especialmente ao meu pai que recentemente
faleceu de COVID-19.

AGRADECIMENTOS

Este trabalho não poderia ser terminado sem a ajuda de diversas pessoas e instituições às quais presto minha homenagem. Certamente estes parágrafos não irão atender a todas as pessoas que fizeram parte dessa importante fase de minha vida. Portanto, desde já peço desculpas àquelas que não estão presentes entre estas palavras, mas elas podem estar certas que fazem parte do meu pensamento e de minha gratidão.

A minha família, pelo carinho, incentivo e total apoio em todos os momentos da minha vida.

Ao meu orientador, que me mostrou os caminhos a serem seguidos e pela confiança depositada.

A todos os professores e colegas da UTFPR, que ajudaram de forma direta e indireta na conclusão deste trabalho.

E por fim, aos companheiros de trabalho na Sanepar, que se dispuseram a me passar informações e acesso aos dados que tornaram possível este trabalho.

Há como imaginar computadores digitais que
fariam bem o jogo da imitação?
(TURING, Alan M., 1950).

RESUMO

NASCIMENTO, Willian Muniz. **Detecção de Vazamentos em Dados de Fluxo de Água com Seleção e Otimização Automática de Modelos**. 2021. 54 f. Dissertação de Mestrado (Mestrado em Computação Aplicada) – Universidade Tecnológica Federal do Paraná. Curitiba, 2021.

O gerenciamento adequado dos recursos hídricos é uma questão prioritária no mundo moderno. Um importante aspecto desta questão é a minimização de perdas na distribuição urbana de água. O monitoramento em tempo real do sistema de distribuição seguido da aplicação de técnicas para detecção de outliers no fluxo de água vem sendo uma alternativa efetiva para a redução desse índice. A identificação dos melhores modelos e parâmetros otimizados para a detecção é um desafio neste cenário complexo. Portanto, a área pode se beneficiar dos desenvolvimentos recentes de estratégias para seleção e ajuste de modelos, área também conhecida por Aprendizagem de Máquina Automatizada, do inglês *Automated Machine Learning* (AutoML). Este trabalho apresenta uma proposta de aplicação de técnicas de detecção de outliers e recursos de AutoML em dados de fluxo de água em 16 Zonas de Pressão do sistema de distribuição de água de Curitiba, estado do Paraná, Brasil. É aplicada uma ferramenta "off-the-shelf"de AutoML e realizada uma otimização automática de algoritmos específicos de detecção de outliers. Os experimentos conduzidos indicam que a combinação de AutoML com técnicas tradicionais de detecção de outliers é o direcionamento mais efetivo.

Palavras-chave: Detecção de Vazamentos. Seleção e Otimização de Modelos. AutoML. Mapas Auto-Organizáveis. Fator de Outlier Local.

ABSTRACT

NASCIMENTO, Willian Muniz. **Leakage Detection using Water Flow Data with Automatic Model Selection and Optimization**. 2021. 54 p. Dissertation (Master's Degree in Applied Computing) – Universidade Tecnológica Federal do Paraná. Curitiba, 2021.

Proper management of water resources is a priority issue in the modern world. An important aspect of this matter is the reduction of losses in the urban water distribution. The real-time monitoring of the distribution system followed by the application of outlier detection techniques on water flow data has been an effective alternative to reduce this index. The identification of best models and optimized parameters for detection is a challenge in this complex scenario. Therefore, the area can benefit from the recent developments in model selection and adjustment strategies, this area also known as Automated Machine Learning (AutoML). This work presents a proposal for the application of outlier detection techniques and AutoML resources on water flow data of 16 District Metering Areas (DMAs) of the water distribution system in Curitiba, state of Paraná, Brazil. An off-the-shelf AutoML tool is applied and automatic optimization of specific outlier detection algorithms is performed. The experiments conducted indicate that combining AutoML with traditional outlier detection techniques is the most effective direction.

Keywords: Leakage Detection. Model Selection and Optimization. AutoML. Self-Organizing Maps. Local Outlier Factor.

LISTA DE ILUSTRAÇÕES

LISTA DE TABELAS

LISTA DE ABREVIATURAS, SIGLAS E ACRÔNIMOS

SIGLAS

AutoML	Aprendizagem de Máquina Automatizada, do inglês *Automated Machine Learning*
LOF	Fator Outlier Local, do inglês *Local Outlier Factor*
PCA	Análise de Componentes Principais, do inglês *Principal Component Analysis*
SCADA	Supervisão e Aquisição de Dados, do inglês *Supervisory Control And Data Acquisition*
SOM	Mapas Auto-Organizáveis, do inglês *Self-Organizing Maps*

ACRÔNIMOS

Z-Score	Pontuação Padrão, do inglês *Standard Score*

SUMÁRIO

1 INTRODUÇÃO

Quando uma companhia de distribuição realiza o abastecimento de água em um município, parte da água não é faturada, ou seja, a quantidade de água captada não é a mesma cobrada dos clientes. Parte dessa água fica no sistema de distribuição caracterizada como perda. O problema das perdas de água tem múltiplas causas: vazamentos, erros de medição, consumos não autorizados, entre outras.

Minimizar o vazamento de água significa ganho de eficiência, visto que será necessário tratar e bombear menos água. Isto se tornará cada vez mais importante devido ao crescente valor da água para a humanidade e as restrições ambientais cada vez maiores.

Perdas de água no sistema de distribuição são atualmente um grande problema para as empresas de saneamento no mundo. Segundo o Instituto Trata Brasil (TRATA BRASIL, 2017) o Brasil apresenta um índice de 38% de perdas de água na distribuição, gerando uma perda financeira acima dos R$ 11 bilhões. Nesta pesquisa utilizaremos como contexto o sistema de distribuição de água de Curitiba, composto por aproximadamente 130 zonas de pressão e que apresenta um índice de 26,16% de perdas de água na distribuição.

Para a identificação de vazamentos, atualmente na cidade de Curitiba são utilizadas planilhas com os valores de fluxo de água agrupados por dia do mês e por zonas de pressão. Os dados são comparados visualmente através de gráficos com foco no uso médio do mês atual e do mês anterior. Estes gráficos mostram os valores de média diária e fluxo mínimo noturno no decorrer do mês, exibindo também os valores do mês anterior.

Identificar vazamentos automaticamente enfrenta alguns desafios, como variação de acordo com o clima, leituras erradas dos sensores, variação devido a feriados, dificuldade de se determinar exatamente o que é um outlier, além da grande quantidade de dados muitas vezes não mapeados. É importante também evitar a geração de falsos positivos, pois ao sinalizar um vazamento é necessário deslocar uma equipe para ir a campo verificar o incidente. Além de causar prejuízos financeiros para a empresa, o excesso de falsos positivos faz o sistema de detecção perder credibilidade, no entanto não é tratado este problema neste trabalho.

Outro desafio neste contexto é a obtenção de dados de treinamento para os algoritmos. Como o registro de vazamentos reais pode não ser feito ou feito de forma indireta, nosso foco é no uso de dados parcialmente correlacionados com os eventos anômalos. A estratégia é, portanto, um tipo de treinamento semi-supervisionado. Esta dissertação apresenta o desenvolvimento de

um sistema de detecção de vazamentos no contexto descrito. A principal contribuição desta pesquisa é a adaptação do problema e algoritmos relacionados para a utilização de técnicas de AutoML. Nossos testes demonstram que nossa solução apresenta benefícios em comparação tanto com implementações tradicionais sem otimização quanto com o emprego de uma solução de AutoML não adaptada.

1.1 OBJETIVO

O objetivo deste trabalho é apresentar uma solução que possa realizar automaticamente a detecção de vazamentos ou eventos anômalos nas zonas de pressão, sem que os especialistas precisem trabalhar com os valores manualmente. Esta solução deve se adaptar a cada zona de pressão utilizando os dados históricos para ajustar a configuração do sistema, exploramos técnicas de otimização automática de hiper-parâmetros.

A solução considera também os seguintes requisitos:

- Treinamento semi-supervisionado das técnicas, usando registros parcialmente correlacionados com anomalias reais (no caso tratado nesta dissertação usamos dados de ocorrências de manutenção);
- Uso de algoritmos específicos para detecção de outliers (diferentes dos algoritmos de classificação tipicamente disponíveis em soluções de AutoML).

1.2 METODOLOGIA

Para demonstrar a eficácia de diferentes níveis de otimização automática de hiper-parâmetros, testamos três abordagens diferentes para o desenvolvimento do sistema de detecção: (i) implementação tradicional, sem otimização automática, (ii) otimização usando uma ferramenta "off-the-shelf" de AutoML, e (iii) otimização automática sobre algoritmos específicos de detecção de outliers.

Inicialmente são utilizadas técnicas de detecção de outlier nos dados de fluxo das zonas de pressão de forma não supervisionada. A primeira é uma técnica baseada em densidade chamada Fator Outlier Local, do inglês *Local Outlier Factor* (LOF), a segunda uma técnica baseada em Rede Neural chamada Mapas Auto-Organizáveis, do inglês *Self-Organizing Maps* (SOM), a terceira uma abordagem estatística chamada Pontuação Padrão, do inglês *Standard*

Score (Z-Score) e por fim um algoritmo especialista (que simula o processo de decisão utilizado atualmente).

Estes algoritmos foram implementados sem o auxílio de ferramentas para a otimização dos modelos. Estas implementações são usadas como linha de base para avaliar as implementações subsequentes, utilizando otimizadores automatizados.

Em seguida, considerando os recentes avanços na área de aprendizagem de máquina, aplicamos a Aprendizagem de Máquina Automatizada (AutoML). O AutoML executa um conjunto de técnicas supervisionadas e as otimiza com os melhores parâmetros, visando obter o melhor modelo para o problema. Por fim, aplicamos técnicas de otimização automática de hiperparâmetros sobre os algoritmos SOM e LOF adaptados para o contexto semi-supervisionado. Diversos parâmetros foram ajustados usando uma implementação de GridSearch.

O restante deste trabalho está organizado da seguinte forma: A Seção 2 apresenta os fundamentos e trabalhos correlatos, com a descrição do problema de perdas de água, técnicas de detecção de outlier e sua aplicação em sistemas de distribuição. A Seção 3 apresenta a origem e tratamento dos dados, bem como uma análise exploratória para auxiliar a compreensão das informações disponíveis. Na Seção 4 descrevemos a aplicação das técnicas sem otimização, a aplicação do AutoML e a otimização de parâmetros das técnicas LOF e SOM. Na Seção 5 discutimos os principais desafios e nuances do contexto desta proposta.

2 FUNDAMENTOS E TRABALHOS CORRELATOS

2.1 PERDAS DE ÁGUA

O índice de perdas de água não faturada se dá pela diferença do total de água captada pelo total cobrado dos clientes. A água não faturada pode ser consequência de fraudes, utilização dos hidrantes, erros de medição ou problemas no sistema de distribuição (TRATA BRASIL, 2017).

As perdas de água nos sistemas de distribuição têm duas categorias: (i) consumo não autorizado e erros de medida, e (ii) perdas físicas (LAMBERT; HIRNER, 2000). As perdas físicas podem ser devido a problemas de vazamento como rompimento de tubulações ou vazamentos nas conexões das tubulações. Os problemas de rompimento das tubulações são geralmente mais fáceis e rápidos de detectar devido ao auxílio da população, que indica os locais dos vazamentos. No entanto, os rompimentos que estão em terrenos arenosos podem não aparecer e levar dias para serem consertados. Por outro lado, os problemas com vazamento em conexões ou pequenas rachaduras podem demorar muito mais tempo para serem detectados.

Na maioria das cidades a rede de distribuição é muito grande para se trabalhar como um todo. Para solucionar essa grande demanda, a rede de distribuição é subdivida em Zonas de Pressão. Uma zona de pressão pode ser uma área específica de abastecimento, um conjunto de residências ou empresas, um bairro, uma vila, etc.

2.2 MONITORAMENTO DE PERDAS

Para tentar reduzir as perdas de água, as empresas de saneamento estão buscando um monitoramento em tempo real da rede e detecção automática de vazamentos, a fim de reduzir o tempo em que um vazamento é detectado e consequentemente consertado.

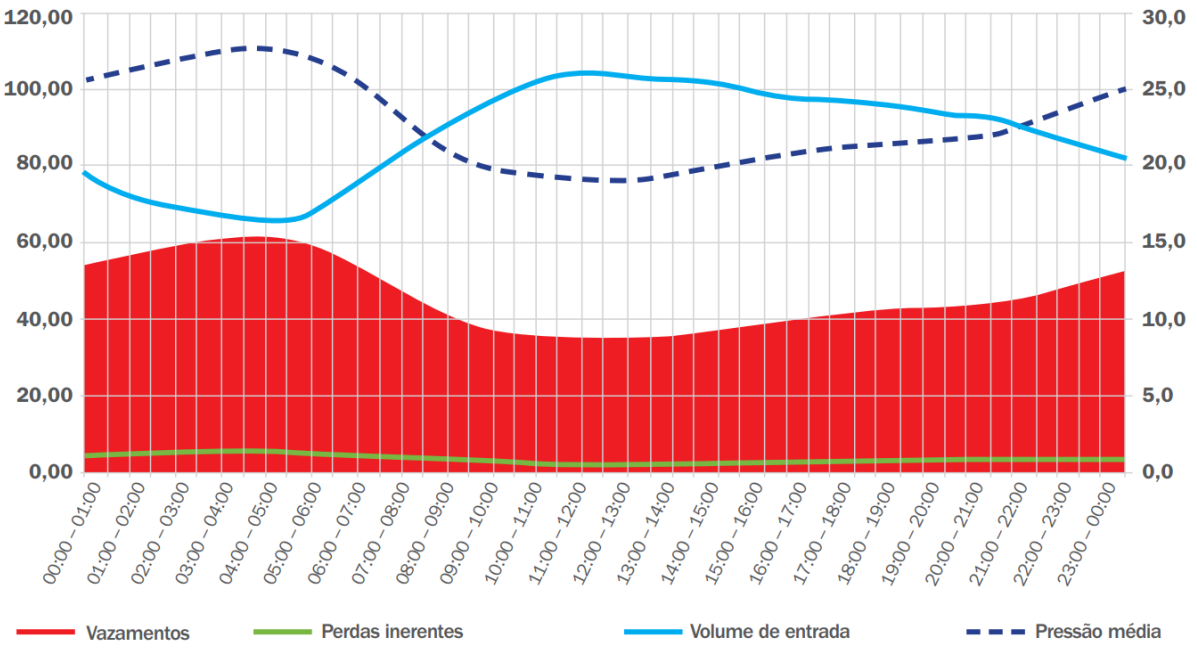
Para realizar o monitoramento são utilizados sistemas de Supervisão e Aquisição de Dados, do inglês *Supervisory Control And Data Acquisition* (SCADA), os Sistemas de Controle Supervisório e Aquisição de Dados monitoram em tempo real muitas Zonas de Pressão ou todas elas. Estes sistemas geram uma grande quantidade de dados, que geralmente são descartados após um certo tempo ou são apenas esquecidos, pois são utilizados somente para monitoramento online.

Na base de dados dos sistemas SCADA é possível encontrar vários tipos de dados operacionais, como por exemplo rotação dos motores, pressão nas tubulações, fluxo de água, porcentagem de válvula aberta, entre outros.

Na análise das zonas de pressão, os dados históricos são muito importantes para distinguir diferentes tipos de outliers e detectar padrões de consumo. Por exemplo, se a zona de pressão é de uma região mais central da cidade, ela geralmente tem o seu uso maior em dias úteis. Zonas de pressão mais afastadas geralmente têm o consumo maior nos finais de semana ou feriados.

Para a detecção de problemas no sistema de distribuição através do fluxo de água, uma medida de suma importância é a vazão mínima noturna, que geralmente utiliza a média ou mediana da vazão de água no período das 00:00 às 05:00 horas. Com essa medida é possível ter um resultado com a mínima interferência do fator humano, e identificar o que é realmente problema no sistema de distribuição. No Gráfico 1 é apresentado o volume de vazamentos, as perdas inerentes, o volume de entrada (vazão) e a pressão média em uma janela de 24 horas de uma zona de pressão. É possível observar a correlação entre vazão, pressão e vazão, no entanto a principal informação está na composição quase que totalmente por vazamento da vazão mínima noturna.

Gráfico 1 – 24 horas de vazamento, vazão e pressão de uma zona de pressão.



Fonte: Guia Prático da AESBE Volume 6

Para identificar os vazamentos são observados os valores de fluxo de água no sistema

de distribuição. As variações nesses valores podem ser devidas a diversos fatores, o que pode dificultar a identificação de um vazamento. Entre os principais fatores estão o uso anormal dos clientes (como em dias de calor com um número maior de banhos, enchimentos de piscinas, feriados e outros), manobras operacionais realizadas na rede de distribuição, problemas com falta de energia, sensores defeituosos, falta de internet e outros (LOUREIRO *et al.*, 2016). Tal variedade dificulta a distinção entre vazamentos reais e desvios justificáveis nos padrões.

2.3 DETECÇÃO DE OUTLIERS

Sistemas de detecção de outliers visam identificar observações que destoam do usual (CHANDOLA *et al.*, 2009). Uma das formas mais simples de se identificar outliers em um conjunto de dados com uma única variável é utilizando um método estatístico chamado *Standard Score* ou Z-Score. Como descrito por Kreyszig (2009), para se calcular o *score* z para o valor x de um conjunto é utilizada a fórmula a seguir.

$$z(x) = \frac{x - \overline{x}}{s} \tag{1}$$

Onde $\overline{x}$ é a média aritmética, e s é o desvio padrão do conjunto. Com a aplicação de um *threshold* sobre os valores de Z-Score obtêm-se quais elementos são outliers. Em geral, as tarefas de detecção de outliers são mais complexas, demandando técnicas mais avançadas.

Detecção de outliers em fluxo temporal é um vasto campo que vem sendo estudado no contexto de aplicação das técnicas a grande volume de dados (GUPTA *et al.*, 2014). Entre os tipos principais de outliers estudados estão os outliers contextuais, que são dependentes e orientados aos tipos dos dados, como dados de séries temporais, dados espaciais, e grafos. Atualmente, problemas de detecção de outliers são tratados majoritariamente como tarefas de aprendizagem de máquina.

Como podemos encontrar em Russell e Norvig (2016), existem quatro tipos de paradigmas de aprendizado que determinam os principais tipos de aprendizagem de máquina: aprendizado não supervisionado, aprendizado por reforço, aprendizado supervisionado e aprendizado semi-supervisionado, como descritos abaixo, baseando-se nas definições de Russell e Norvig (2016).

Em aprendizado não supervisionado o agente aprende padrões nas entradas mesmo que não seja fornecido um feedback. A tarefa mais comum de aprendizado não supervisionado é o

agrupamento: detectando potenciais grupos úteis dos exemplos de entrada.

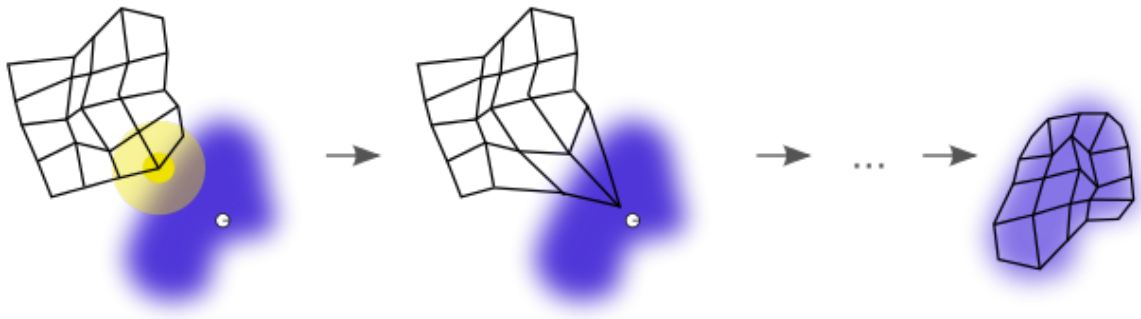
Em aprendizagem de reforço o agente aprende a partir de punições ou recompensas recebidas, punições para o agente identificar que fez algo errado e recompensas para identificar que fez algo correto.

Em aprendizagem supervisionada o agente observa alguns exemplos de pares de entrada e saída, então aprende uma função que mapeia da entrada até a saída. Dadas as entradas e saídas, geralmente como entrada um vetor e como saída um rótulo, é realizado o mapeamento.

O aprendizado semi–supervisionado é um ponto entre o não supervisionado e o supervisionado, isto pode ocorrer devido à falta de mapeamento das entradas e saídas, ou no caso de ruídos nos dados, dados incorretos que levam a um falso mapeamento. Sendo assim necessário fazer a melhor predição possível com os pares de entrada e saída disponíveis.

Nos problemas de detecção de vazamento de água, em geral não há dados confiáveis de treinamento. Portanto, técnicas não supervisionadas ou semi-supervisionadas tendem a ser favorecidas.

Figura 1 – Ilustração do treinamento do SOM.



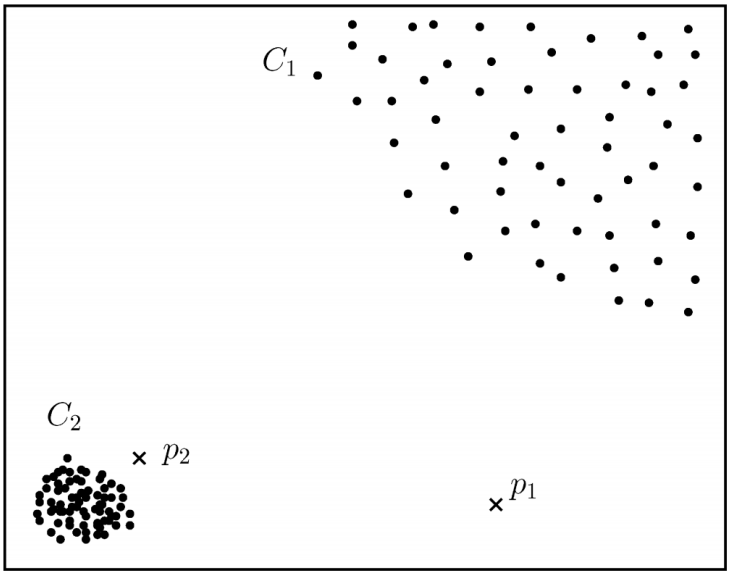
Fonte: Wikimedia Commons

No conjunto de técnicas não supervisionadas utilizadas em tarefas de detecção de outliers podemos citar os Mapas Auto-Organizáveis (Self-Organizing Maps) (KOHONEN; HONKELA, 2007) que são um tipo de rede neural artificial utilizada para análise e visualização de dados em múltiplas dimensões. Os SOMs consistem em uma grade de aprendizagem competitiva que se adapta aos dados. Diferentemente das outas redes neurais que funcionam por correção de erros, nos mapas auto-organizáveis são calculadas as distâncias euclidianas para todos os itens utilizando pesos. Através de um aprendizado competitivo com ajuste dos pesos, o mapa se reestrutura de acordo com a distribuição dos dados no conjunto de entrada. Conforme ilustrado na Figura 1, em azul está a distribuição dos dados de treinamento, o disco branco é o item inicial selecionado dos dados de treinamento, na primeira imagem o nó da grande mais próximo do item

é selecionado em amarelo, seus pesos são atualizados como na segunda imagem, este processo é repetido muitas vezes e a grade final fica reestruturada como na imagem da direita.

As técnicas de detecção de anomalia baseadas em densidade estimam a densidade da vizinhança de cada instância de dado (CHANDOLA *et al.*, 2009). A instância que está em uma vizinhança com baixa densidade é declarada como uma anomalia, enquanto uma instância que está em uma vizinhança com alta densidade é declarada como normal. Estas técnicas têm desempenho reduzido caso os dados tenham regiões de densidade variadas. Para lidar com o problema de variação das densidades nos conjuntos de dados, algumas técnicas foram propostas para computar a densidade das instâncias relativa à densidade de seus vizinhos.

Figura 2 – Detecção de outlier em densidade relativa.



Fonte: (CHANDOLA *et al.*, 2009).

Local Outlier Factor (LOF) (CHANDOLA *et al.*, 2009) é um exemplo de técnica baseada em densidade também utilizada em tarefas de detecção de outliers. Para cada instância de conjunto de dados, o valor de LOF é dado pela proporção da densidade média local dos k-vizinhos mais próximos pela densidade local da instância.

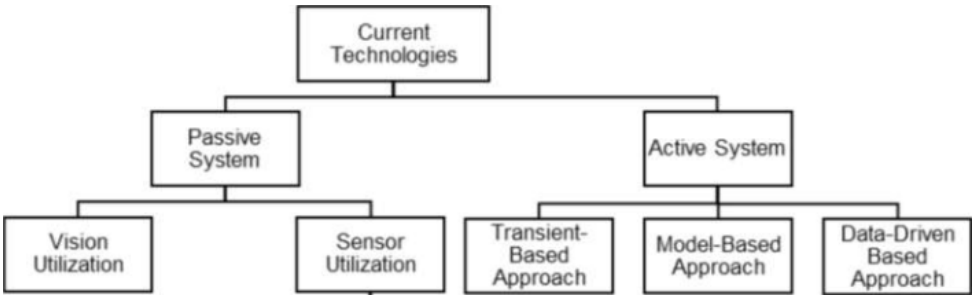
Podemos observar um exemplo prático na Figura 2, onde há dois conjuntos C1 e C2 e dois pontos para serem identificados como outlier p1 e p2. Com densidades diferentes, o conjunto C1 é mais esparso do que o C2, causando problemas para técnicas de detecção baseada em densidade identificarem o p2 como outlier, pois a distância do p2 com seus vizinhos pode ser confundida com as distâncias do conjunto C1. Desta forma, aplicando uma técnica de densidade relativa como o LOF é possível identificar ambos p1 e p2 como outliers, visto que é utilizada a densidade da vizinhança.

2.4 DETECÇÃO DE OUTLIERS EM DISTRIBUIÇÃO DE ÁGUA

A utilização de tecnologia na área de detecção de vazamento em sistemas de distribuição de água vem sendo amplamente estudada, conforme podemos observar nas revisões (LI *et al.*, 2015; PUUST *et al.*, 2010; WU; LIU, 2017; CHAN *et al.*, 2018). As duas principais classificações das soluções propostas para este problema são: Avaliação de Vazamentos e Detecção de Vazamentos (também chamada de Modelo de Controle de Vazamentos em (PUUST *et al.*, 2010)). O objetivo da Avaliação de Vazamentos é estimar a quantidade de perdas de água no sistema de distribuição. Na categoria de Detecção de Vazamentos, o objetivo é detectar e localizar os vazamentos, seguindo dois modelos principais: ativo e passivo. No modelo passivo temos a utilização da visão e utilização de sensores. No modelo ativo temos abordagem baseada em transiente, abordagem baseada em modelo e abordagem orientada aos dados.

Conforme apresentado em Chan *et al.* (2018) uma visão global dos diferentes tipos de tecnologias disponíveis atualmente para detecção de vazamentos pode ser vista na Figura 3. Estas tecnologias são descritas a seguir.

Figura 3 – Visão geral das tecnologias atuais de detecção de vazamentos.



Fonte: Figura adaptada de (CHAN *et al.*, 2018).

Modelo de Utilização da Visão (Vision Utilization): De maneira visual observar o chão em busca de vestígios de vazamento como degradação do solo ou vegetação crescendo de forma anômala indicando possível vazamento.

Modelo de Utilização de Sensores (Sensor Utilization): Os dispositivos para detectar vazamento foram evoluindo com a tecnologia, um dos mais clássicos sensores e ainda amplamente utilizado é o geofone portátil. Nesta abordagem, os técnicos vão passando com o equipamento nas ruas por cima da tubulação e com ouvidos treinados detectam os vazamentos.

Abordagem Baseada em Transiente (Transient-Based Approach): Um vazamento é um fenômeno hidráulico. Desta forma o vazamento pode ser detectado através do transiente hidráulico. O transiente hidráulico é uma onda de pressão de curta duração. No momento em que

ocorrer o início de um vazamento há uma queda na pressão que produz o fenômeno de transiente hidráulico. O método de detecção baseado nesse fenômeno consiste em medir e modelar os traços do transiente hidráulico na rede de tubulação. Esta técnica geralmente necessita de uma grande quantidade de sensores.

Abordagem baseada em Modelo (Model-Based Approach): A abordagem baseada em modelos geralmente envolve a utilização funções ou fórmulas matemáticas que representem a operação do sistema de distribuição estudado. É possível com essa abordagem detectar a localização do vazamento comparando a pressão medida com a predição do modelo. Contudo é necessário que o modelo tenha uma boa representação do sistema.

Abordagem Orientada aos Dados (Data-Driven Approach): A abordagem orientada a dados depende da coleta dos dados através dos sensores, do processamento desse sinal e de uma análise estatística para ser utilizada como detecção de vazamento. A vantagem dessa abordagem é não precisar de um conhecimento aprofundado do sistema de distribuição de água, pois basta aprender do histórico de dados coletados utilizando ferramentas estatísticas ou reconhecimento de padrões. E a principal desvantagem dessa abordagem é a grande quantidade de dados necessária para criar um modelo preditivo ou de classificação. Esta será a abordagem utilizada nesta pesquisa.

Conforme apresentado por Wu e Liu (2017), é possível encontrar diversas técnicas para identificar vazamentos utilizando abordagens orientada a dados. Elas podem ser separadas em três categorias: método de classificação, método de predição-classificação, e método estatístico.

Um método de classificação pode ser construído para distinguir um vazamento dos dados normais. No estudo Aksela *et al.* (2009) os autores apresentaram um método para detecção de vazamentos em sistemas de distribuição de água baseado em Self Organazing Maps (SOM). Os autores obtêm os dados de três medidores de vazão, com leituras em formato de média horária, na rede de distribuição estudada e têm a localização física de oito vazamentos. Com isso, eles criaram uma função de vazamento com base nestes dados e princípios hidráulicos para facilitar o treinamento do SOM.

Os autores Aksela *et al.* (2009) utilizaram uma estrutura hexagonal para evitar favoritismo das direções horizontais ou verticais e um formato de folha para o mapa. No treinamento foram utilizados os dados dos três medidores de vazão por dia da semana como vetores de entrada (21 dimensões). Os valores da função de vazamento também são adicionados no processo de treinamento. O resultado mostra que o modelo treinado detecta os vazamentos em uma

determinada área da rede de distribuição. O método apresentado nesse trabalho opera sozinho após o treinamento do SOM. Informações sobre vazamentos são necessárias somente na fase de treinamento. Por utilizar esses dados de vazamentos no treinamento, o trabalhos desses autores está na categoria dos métodos de classificação.

O modelo apresentado por Aksela *et al.* (2009) é dependente do conjunto de treinamento utilizado (dos vazamentos indicados previamente), podendo enviesar o algoritmo caso algum vazamento não seja marcado ou caso seja marcado um vazamento que não ocorreu. Porém, a abordagem se comportou bem no cenário real aplicado.

Os autores Romano *et al.* (2010) apresentaram uma metodologia online para detecção automática de vazamentos de pequeno e grande porte analisando os dados coletados de sensores. O estudo de caso utilizou uma *District Metering Area* (DMA ou Zona de Pressão) no Reino Unido. A validação da metodologia se deu por simulação de eventos de vazamentos. Esta metodologia utilizou várias técnicas de inteligência artificial, *wavelets* para tirar o ruído dos dados de fluxo e pressão, redes neurais para previsão de curto prazo dos valores de pressão e fluxo, controle estatístico de processos para analisar as discrepâncias entre os dados de previsão e os dados observados, e por fim, realizou uma classificação destas discrepâncias e alarmes através de um sistema de inferências baseado em redes *bayesianas*. Os resultados se mostraram satisfatórios com nenhum falso alarme. E com intervalo de execução de 15 minutos o sistema conseguiu detectar todos os vazamentos simulados.

Os mesmos autores têm uma segunda publicação (ROMANO *et al.*, 2012) mostrando algumas evoluções na metodologia, utilizaram o controle estatístico de processos para analisar tantos os vazamentos de curto quanto os de longo prazo. Para validação da metodologia os autores utilizaram várias DMAs no Reino Unido com eventos anômalos reais e simulados.

Para detectar eventos anômalos no fluxo de água de cinco zonas de pressão, os autores Loureiro *et al.* (2016) criaram uma abordagem prática utilizando dados históricos do sistema SCADA com frequência de até 15 minutos e um histórico de ordens de serviço. Os autores propuseram quatro novos métodos estatísticos para detecção de eventos anômalos baseados em regiões de outlier, nos quais há um limite inferior e um limite superior dos valores de fluxo. Caso o valor de entrada esteja fora dessa região ele é considerado um outlier. Esta região foi determinada previamente por um conjunto de treinamento no qual os dados estavam marcados como sendo outlier ou não. Para realizar esta marcação os autores utilizaram ordens de serviços. As ordens de serviços podem conter informações incompletas, ou simplesmente não ter sido

registrado uma ordem, portanto, estas não representavam especificamente vazamentos (similar aos dados de manutenção que utilizamos neste trabalho). Os autores utilizaram curvas ROC para fazer comparações entre os métodos treinados com duas janelas de tempo diferentes. Por fim, foram selecionadas abordagens para identificar dois tipos de eventos anômalos: eventos instantâneos de anomalia ou eventos relacionados a vazão mínima noturna (entre as horas 01:00 e 06:00). Os autores obtiveram resultados satisfatórios nas zonas de pressão majoritariamente residenciais para detectar grandes vazamentos e outros eventos similares.

Palau *et al.* (2011) utilizaram para aquisição dos dados um sistema SCADA e aplicaram a sua metodologia no nível de DMA. Os autores aplicaram métodos estatísticos multivariados no controle de entrada de água nas DMAs. A técnica aplicada foi a Análise de Componentes Principais, do inglês *Principal Component Analysis* (PCA), utilizada para simplificar os dados de vazão adquiridos através dos sistemas SCADA, sintetizando as informações em um modelo estatístico capaz de explicar o comportamento da rede de distribuição de água. Através da técnica foi possível gerar gráficos de controle para os operadores identificarem anomalias no uso da água, vazamentos ou conexões ilegais. A base da metodologia apresentada é a entrada de fluxo de água na DMA. A sensibilidade da técnica depende fortemente da qualidade e variabilidade dos dados usados para construir o modelo. Para isso os autores extraíram as entradas de fluxo diariamente em dias úteis, finais de semana e intervalos de horas diferentes. Assim a variação dos dados é reduzida consideravelmente e a efetividade da detecção de vazamento é alcançada. Entre os dois recursos estatísticos (T^2 *hotelling* e DMOD) para detecção de outliers, o DMOD mostrou melhor sensibilidade na detecção de grandes vazamentos. Os autores provaram que a técnica tem alta capacidade de detecção de grandes vazamentos e consumos inesperados, um vazamento de 5% da média do fluxo pode ser detectado com uma probabilidade de 30% a 95%, dependendo da hora da ocorrência, a maior eficiência ocorre durante a madrugada.

O trabalho Mounce *et al.* (2003) apresenta análise e fusão de dados para dados de pressão e fluxo em um sistema de água tratada. Para predição e classificação dos vazamentos foram utilizadas redes neurais artificias. Um sistema baseado em regras realiza uma fusão da saída da rede neural para produzir uma classificação geral para um conjunto de zonas. Os resultados são de um local experimental da rede de distribuição de uma companhia de água do Reino Unido. Os vazamentos foram simulados abrindo hidrantes. O sistema conseguiu detectar os eventos adequadamente.

Os trabalhos citados na Tabela 1 têm algumas características em comum e utilizam

Tabela 1 – Comparativo entre os trabalhos encontrados na literatura.

Trabalho	**Contexto**	**Dados de entrada**	**Elementos contextuais**	**Abordagem**	**Método de avaliação**
(AKSELA *et al.*, 2009)	Helsinki.	Fluxo de água.	Média diária dos sete dias da semana.	Função de Vazamento; SOM.	Erro Quadrático Médio.
(LOUREIRO *et al.*, 2016)	5 DMAs de Lisboa.	Fluxo de água.	Fluxo mínimo noturno; Dias da semana; Sábados.	Método estatístico.	Curvas ROC usando dados de Ordens de Serviço.
(MOUNCE *et al.*, 2003)	15 DMAs.	Fluxo de água; Pressão; Sensores de Opacidade.	Dia do mês; Fluxo mínimo noturno.	Rede Neural.	Vazamentos simulados.
(PALAU *et al.*, 2011)	DMA da Espanha.	Fluxo de entrada de água na DMA.	Dia do Mês; Dias úteis; Finais de semana Períodos (manhã, tarde, noite); Fluxo Noturno.	PCA.	Eficácia; Nível de confiança.
(ROMANO *et al.*, 2010)	DMA do Reino Unido.	Fluxo de água; Pressão.	Data e Hora.	Wavelets; Rede Neural Artificial; Rede Bayesiana.	Vazamentos simulados.
Este trabalho	15 DMAs de Curitiba.	Fluxo de água.	Dia do mês; Dia da semana; Fluxo mínimo noturno; Média diária.	AutoML; SOM; LOF.	F-Score; Dados de manutenção.

Fonte: Autoria própria.

técnicas diferentes para tratar o problema de detecção de vazamento de água. As principais semelhanças entre este trabalho e os encontrados na literatura são a utilização do contexto de zonas de pressão (DMA), o fluxo de água como dado de entrada e os elementos contextuais como: dias do mês, dias da semana e fluxo mínimo noturno. Algumas características deste trabalho se diferem dos outros encontrados. Incluímos elementos contextuais não utilizados nos outros trabalhos. E visando representar as dificuldades reais do problema, neste trabalho é utilizada uma quantidade maior de zonas de pressão do que a maioria dos trabalhos encontrados, além de utilizarmos tanto zonas de pressão de recalque quanto por gravidade. As principais diferenças, no entanto, se dão pela utilização de AutoML para fazer a seleção do modelo e otimizar os hiper-parâmetros, além da aplicação semi-supervisionada do SOM e LOF com otimização dos hiper-parâmetros para cada zona de pressão. Na seção 4 são descritas com mais detalhes as características da implementação deste trabalho.

2.5 DETECÇÃO DE OUTLIERS EM PROBLEMAS SIMILARES

Nesta seção apresentamos duas pesquisas de problemas que apresentam características parecidas com a detecção de vazamentos.

O autor Junior (2018) apresentou uma metodologia de análise para identificar possíveis

anomalias de consumo de energia, e desta forma reduzir o número de fraudes no sistema de medição. A pesquisa utilizou uma técnica não supervisionada, Mapas Auto-Organizáveis (SOM). Para apresentar o melhor resultado na sensibilidade da rede em identificar alterações no padrão de consumo foram utilizadas diferentes grandezas elétricas a aplicação de uma normalização por tangente hiperbólica modificada, pois normalmente esse método é utilizado para evitar a saturação dos neurônios interconectados nas redes perceptron. O processo de validação foi através de simulação dos valores da corrente elétrica, analisando o consumidor individualmente sem formar grupos. Foram utilizadas medições dos parâmetros elétricos armazenados pelos medidores eletrônicos de energia. No que se refere a consumo de água, uma fraude pode ser considerada um evento anômalo, levando as técnicas de detecção de outliers a identificarem estes eventos como vazamentos.

No trabalho Benghi (2020) é apresentado um modelo para detecção e interpretação de falhas em locomotivas. A proposta busca não somente realizar o reconhecimento da anomalia como também apresentar uma explicação sobre o diagnóstico utilizando *Outlier Aspect Mining* (OAM). Inicialmente foi gerado um modelo para detecção de anomalias a partir de dados históricos e posteriormente foi aperfeiçoado para diagnóstico em tempo real. A abordagem proposta consiste em identificar as principais falhas ocorridas em tempo real. A abordagem é dependente da atualização de novos dados para manter o desempenho constante. Sobre os outliers detectados foram utilizados os métodos de OAM para a contextualização e interpretação das falhas. A detecção de falhas em locomotivas se assemelha com este trabalho por utilizar técnicas não supervisionadas e dados históricos para detecção de outliers. A aplicação em tempo real e a interpretação dos outliers são características desejadas em detecção de vazamentos.

2.6 APRENDIZAGEM DE MÁQUINA AUTOMATIZADA

Um dos problemas encontrados na aplicação de algoritmos de detecção de outlier em sistemas de distribuição de água é a complexidade para encontrar os melhores modelos, *features* e parâmetros. Atualmente uma solução para esse tipo de problema é a utilização da Aprendizagem de Máquina Automatizada (AutoML). AutoML envolve diversos aspectos: Engenharia de Recursos (*Feature engineering*), Combinação de Seleção de Modelo e Otimização de Hiper-parâmetro (*Combined Model Selection and Hyperparameter optimization*), e Meta-Aprendizagem (*Meta-Learning*). Os conceitos envolvidos são descritos a seguir.

Engenharia de recursos tem como objetivo apresentar as melhores *features* para deter-

minado modelo. Para realizar isto é necessário o pré-processamento, aprendizagem da representatividade e então a seleção das principais *features* (TUGGENER *et al.*, 2019).

Entre as opções para implementação da engenharia de recursos temos um algoritmo de aprendizagem de *features* baseado em regressão chamado AutoLearn (KAUL *et al.*, 2017) e um framework para geração automatizada de *features* chamado ExploreKit (KATZ *et al.*, 2016). Ambas opções têm um objetivo parecido: aliviar o trabalho dos especialistas na engenharia de recursos através da automatização deste processo. Desta forma as duas soluções podem fazer a seleção das melhores *features* e também apresentar *features* candidatas tanto modificando as existentes quanto gerando novas.

Hutter *et al.* (2019) descrevem que os hiper-parâmetros podem ser encontrados na maioria dos algoritmos de aprendizagem de máquina, especialmente nos algoritmos mais recentes de *deep learning*. Estes hiper-parâmetros são configurados nos algoritmos antes da execução e não sofrem alteração durante o processo de treinamento do algoritmo. Dessa forma a automatização da otimização de hiper-parâmetros se torna uma tarefa essencial para reduzir tempo necessário de especialistas para configurá-los.

Conforme apresentado por Yu e Zhu (2020) as automatizações da otimização dos hiper-parâmetros são geralmente realizadas por algoritmos de busca como: *Grid Search*, *Random Search*, *Bayesian Optimization* e seus variantes, e *Tree Parzen Estimators*. Os algoritmos de busca são implementados e estão disponíveis para utilização em kits de ferramentas que se propõem a solucionar esse problema. Os autores também apresentam alguns kits de ferramentas: Google Vizier, Amazon SageMaker, Neural Network Intelligence by Microsoft, e Ray.Tune. Estas ferramentas são divididas em duas categorias, as de código aberto e as de computação em nuvem. As ferramentas que necessitam de recursos de computação em nuvem da Google e Amazon são pagas e o código dos algoritmos é fechado.

A princípio todo serviço de aprendizagem de máquina a ser aplicado em um conjunto de dados precisa definir qual algoritmo vai utilizar e como configurar seus hiper-parâmetros, além de saber como e quando processar suas *features* (FEURER *et al.*, 2015). A combinação de seleção automática de algoritmo e hiper-parâmetros foi definida como um problema por Thornton *et al.* (2013), chamado problema CASH (*Combined Algorithm Selection and Hyperparameter optimization*). Este problema consiste em selecionar automaticamente o algoritmo de aprendizado e simultaneamente seus hiper-parâmetros de forma a otimizar o desempenho, dado um conjunto de dados.

Uma proposta para resolver o problema CASH é através da utilização dos recursos de aprendizagem de máquina como se fossem blocos montados em um pipeline. Um pipeline completo de aprendizagem de máquina consiste nos blocos: limpeza dos dados, engenharia de recursos, seleção de modelo, otimização de hiper-parâmetro e ao final criar um ensemble dos principais modelos treinados para obter um bom desempenho nos dados de teste (TUGGENER *et al.*, 2019).

A fim de automatizar a biblioteca de aprendizagem de máquina em Python chamada scikit-learn, (FEURER *et al.*, 2015) apresentaram seu sistema Auto-sklearn. Esse sistema realiza aprendizagem de máquina automatizada baseada na definição de problema CASH do Auto-WEKA. Os autores propuseram uma ferramenta robusta e eficiente para solução desse problema. O pipeline de aprendizagem de máquina do sistema Auto-sklearn é dado em três blocos: Meta-aprendizagem, Otimizador Bayesiano e Construtor de Ensemble. Primeiro é utilizada a Meta-aprendizagem para buscar no conjunto de dados os frameworks de aprendizagem com bom desempenho para partir de um bom ponto inicial no Otimizador Bayesiano, segundo é construído automaticamente pelo Otimizador Bayesiano o conjunto de modelos e por fim é projetada cuidadosamente a estrutura de aprendizagem de máquina altamente parametrizada a partir de classificadores e pré-processadores de alto desempenho da biblioteca scikit-learn.

Esta pesquisa empregou tecnologias de AutoML no problema de detecção de outliers em sistemas de distribuição de água. Neste trabalho, o foco é em seleção de modelos e otimização de hiper-parâmetros. Uma das nossas contribuições é a adaptação dos algoritmos não supervisionados SOM e LOF para um esquema semi-supervisionado com otimização automática de hiper-parâmetros.

3 DADOS E ANÁLISE EXPLORATÓRIA

Nessa seção apresentamos detalhes sobre a origem dos dados da base temporal, o tratamento realizado para a base relacional, a estratégia de obtenção dos dados que indicam eventos anômalos e uma análise exploratória dos dados mais importantes.

3.1 ORIGEM E TRATAMENTO DOS DADOS

Utilizando o sistema de distribuição de água de Curitiba, foram definidas 16 zonas de pressão em conjunto com os especialistas do Centro de Controle e Operação (CCO) para realizar esta pesquisa. Os dados foram obtidos de uma base de dados temporal chamada Proficy Historian[1]. Foi necessário criar um mapeamento em base relacional das entradas e saídas de cada Zona de Pressão e também uma limpeza dos dados. O período utilizado para a aplicação das técnicas foi de setembro a dezembro de 2018.

3.1.1 Base de Dados Temporal

Cada tipo de equipamento (bomba, sensor, válvula, etc.) produz um conjunto de variáveis numéricas diferentes que são armazenadas na base Historian. Estas variáveis podem ser referentes aos valores de: velocidade de rotação, tensão elétrica, fluxo de água, tempo em execução, pressão, entre outros. Esses valores numéricos são armazenados em uma base de dados temporal com a denominação genérica de *"tags"*. Cada informação numérica é registrada em uma *tag* com o seu devido componente temporal.

A base de dados temporal Historian, é gerada pelo sistema SCADA de monitoramento e controle em tempo real. Esta base é composta por *tags* com as seguintes colunas: Tag Name, Historian Tag Name, TimeStamp, Value, Quality. A base de dados é composta por aproximadamente dez mil *tags*, com diversos valores armazenados: valores de fluxo, pressão, tensão, frequência, tempo de execução, etc. Cada uma dessas métricas é um valor decimal em série temporal, gravado em uma *tag*, independente das demais. Ou seja, para cada leitura é gravado um valor e uma data/hora, além das outras colunas da *tag*. Devido ao fato de não haver relacionamento entre as *tags*, esta base utiliza um padrão de nomenclatura que torna possível a identificação de algumas

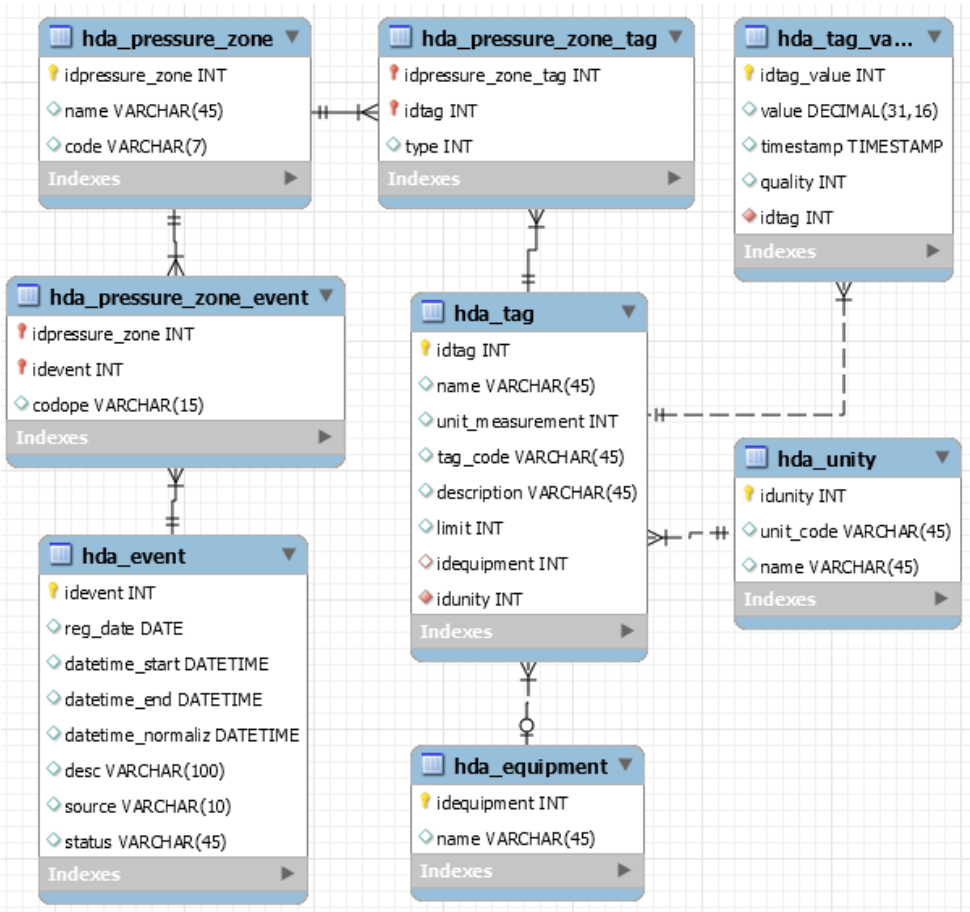
[1] https://www.ge.com/digital/applications/proficy-historian

informações: a unidade, o tipo, e o código do equipamento. Contudo estas informações seriam insuficientes para realizar essa pesquisa. Desta forma foi importado para uma base de dados relacional um mapeamento entre as *tags* de fluxo de água e as Zonas de Pressão.

3.1.2 Base de Dados Relacional

Os dados da base de dados temporal foram mapeados e importados para uma base de dados relacional conforme modelo na Figura 4.

Figura 4 – Modelo do banco de dados relacional contendo as tabelas que integram as informações usadas neste trabalho.



Fonte: Autoria Própria.

Na Figura 4, uma Zona de Pressão (*pressure_zone*) é uma região específica da rede de distribuição, podendo ser um bairro, vila, ou apenas um conjunto de ligações da rede. Um equipamento (*equipment*) pode ser um sensor para medir fluxo, volume, pressão, ou até mesmo um motor. Uma Unidade (*unity*) é uma divisão mais abrangente pode ser um município, região metropolitana, conjunto de municípios. Uma *tag* é uma medida única do equipamento como litros por segundo, metros cúbicos, partes por metro, etc. Os Valores das *tags* (*tag_value*) são

valores numéricos em datas com qualidade boa ou ruim. Um Evento (*event*) é um registro de manutenção na rede de distribuição de água.

As Zonas de Pressão se relacionam com as *tags* de maneira que uma Zona de Pressão pode ser composta por várias *tags*, algumas de entrada, outras de saída. Uma *tag* de entrada de uma Zona de Pressão pode ser uma *tag* de saída de outra Zona de Pressão, ou seja, uma *tag* pode estar ligada a várias Zonas de Pressão.

As Zonas de Pressão passam por manutenções, algumas situações que interrompem o abastecimento e/ou afetam os dados. Estas manutenções são chamadas de Eventos (*event*) e importadas com uma data de início, data de fim e data de normalização do incidente. Um evento pode ser um conserto na rede, uma rede rompida, um problema de automação, o registro de uma queda de energia, entre outros. Os dados utilizados como eventos foram obtidos e mapeados de outro sistema. Estes dados não representam um vazamento, no entanto são considerados anomalias (*outliers*) neste trabalho, por representarem situações reais que ocorrem na rede de distribuição. Desta forma os eventos servem para realizar uma comparação entre as técnicas de detecção de anomalias na seção 4.

Um Evento (*event*) é ligado a uma ou mais Zonas de Pressão através de um código CODOPE. O código CODOPE é a identificação de uma região específica, através dele é possível identificar a Zona de Pressão, Unidade e outras informações.

3.1.3 Obtenção de Dados de Possíveis Eventos Anômalos

Atualmente em Curitiba há dois sistemas onde são informados dados de manobras e ocorrências na rede de distribuição. O primeiro é um sistema em plataforma mainframe que contém dados de ocorrências, manobras e manutenções. O segundo é um sistema em plataforma baixa que contém dados de ocorrências e manutenções. Uma ocorrência pode ser uma tubulação rompida, um hidrômetro que parou de funcionar, uma falha na rede de comunicação, etc. Uma manobra é definida quando é necessário o desvio de água de uma região para outra, seja por elevatórias com utilização de bombas ou por gravidade. Uma manutenção pode ser a troca de peças com defeito, substituição de equipamentos, limpeza, etc.

Estes dados de ocorrências, manobras e manutenções de ambos os sistemas podem não ser eventos anômalos, visto que nem todos estes tipos de serviços causam um impacto no sistema de distribuição.

Nesta pesquisa, utilizamos os dados do sistema em plataforma mainframe devido ao

fato dele conter em seus lançamentos os códigos CODOPEs, nos quais é possível obter os códigos das zonas de pressão. O sistema web contém dados importantes para a validação dos algoritmos e que impactam na qualidade das métricas aplicadas para avaliação da qualidade dos resultados. No entanto este sistema não está sendo utilizado devido ao fato dos dados estarem ligados às unidades, portanto não contêm uma ligação direta com as zonas de pressão e não temos um mapeamento no momento de quais zonas de pressão pertencem a quais unidades. Isto está inclusive previsto na base de dados relacional, mas ainda não está disponível.

Para a importação destes dados do sistema em plataforma mainframe de manobras e manutenções para a base de dados relacional foi necessária uma limpeza nos dados, excluindo os que não continham data, os que não continham o código CODOPE, e os que não estavam na situação finalizados. Estes dados foram chamados na base relacional de eventos. Cada evento pode conter um ou mais CODOPE, consequentemente pode estar ligado a uma ou mais zona de pressão. Os dados de manutenção são utilizados como dados de treinamento semi-supervisionado e como referência para a avaliação da qualidade dos algoritmos (seção 4).

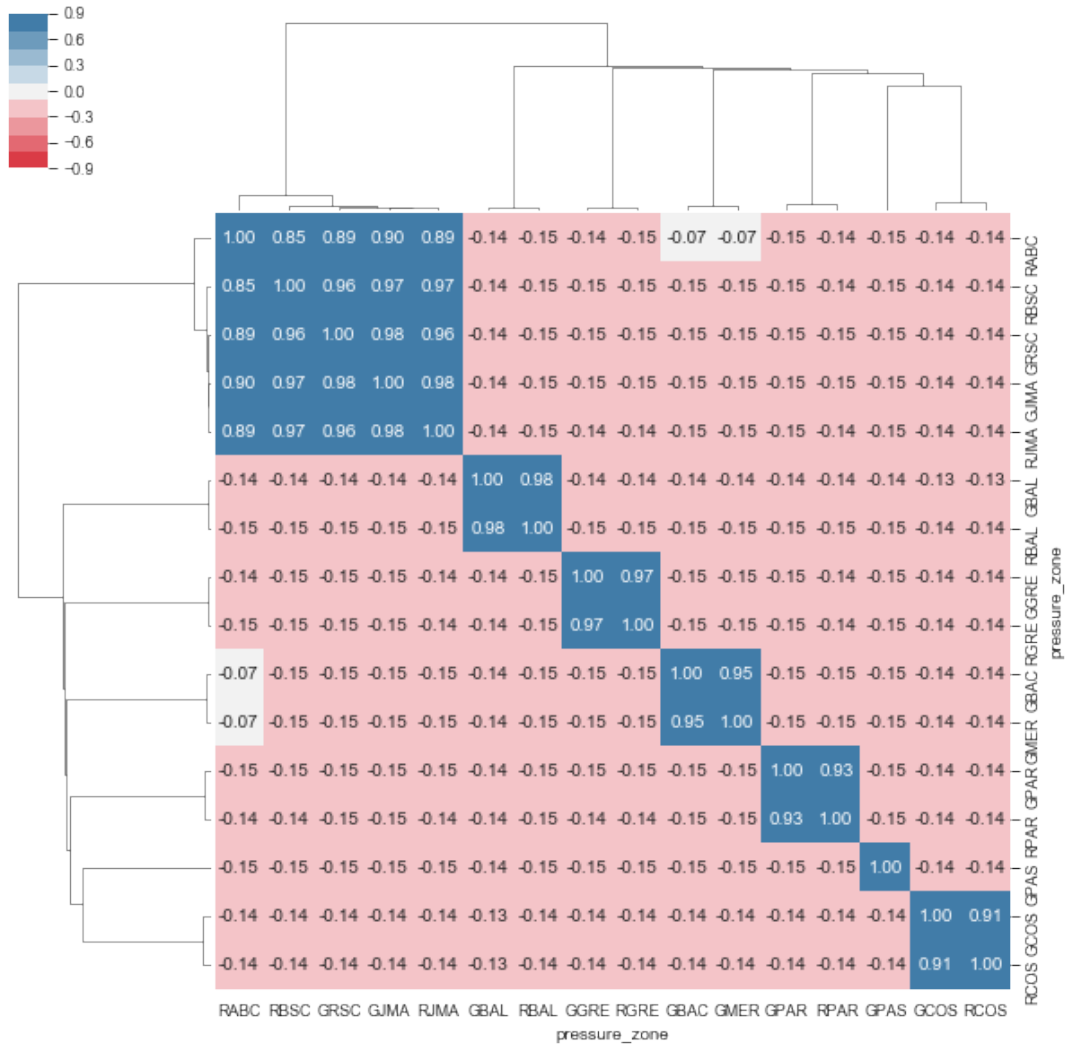
3.2 ANÁLISE EXPLORATÓRIA

Com os dados importados para a base de dados relacional foi realizada uma análise exploratória para identificar as principais características dos dados. Os nomes das dezesseis zonas de pressão utilizadas foram transformadas em siglas para ofuscar a localização, contudo a primeira letra é R para recalque e G para gravidade. Recalque é quando a zona de pressão necessita de bombeamento, e a outra é por gravidade.

Na sequência de gráficos a seguir foram utilizados dados de mínima e média de fluxo de quatro meses, de setembro a dezembro de 2018. Nas correlações mostradas no Gráfico 2, algumas zonas de pressão têm uma alta correlação. Esta alta correlação ocorre quando as zonas de pressão são semelhantes em suas características físicas. Por exemplo, estarem localizadas fisicamente próximas, ou no caso de apresentarem um conjunto de ligações majoritariamente residenciais ou comerciais, com isso dispondo de um padrão de consumo parecido.

O histograma apresentado no Gráfico 3 é de uma zona de pressão como amostra da distribuição dos valores de média e mínima das zonas de pressão. Os valores de mínima e média tendem a ficar separados devido ao fato da mínima ser um horário de menor utilização. Todas as zonas de pressão apresentam uma certa variação dentro da média e mínima, no entanto essa variação tende a seguir um padrão.

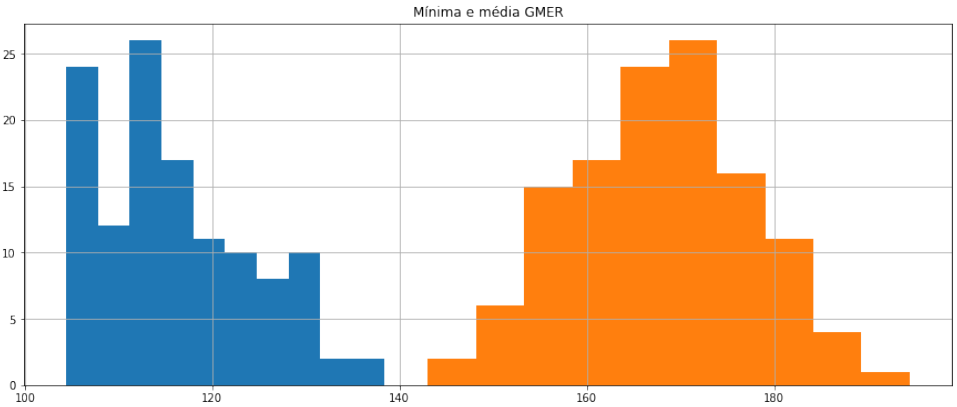
Gráfico 2 – Correlação entre as Zonas de Pressão.



Fonte: Autoria própria.

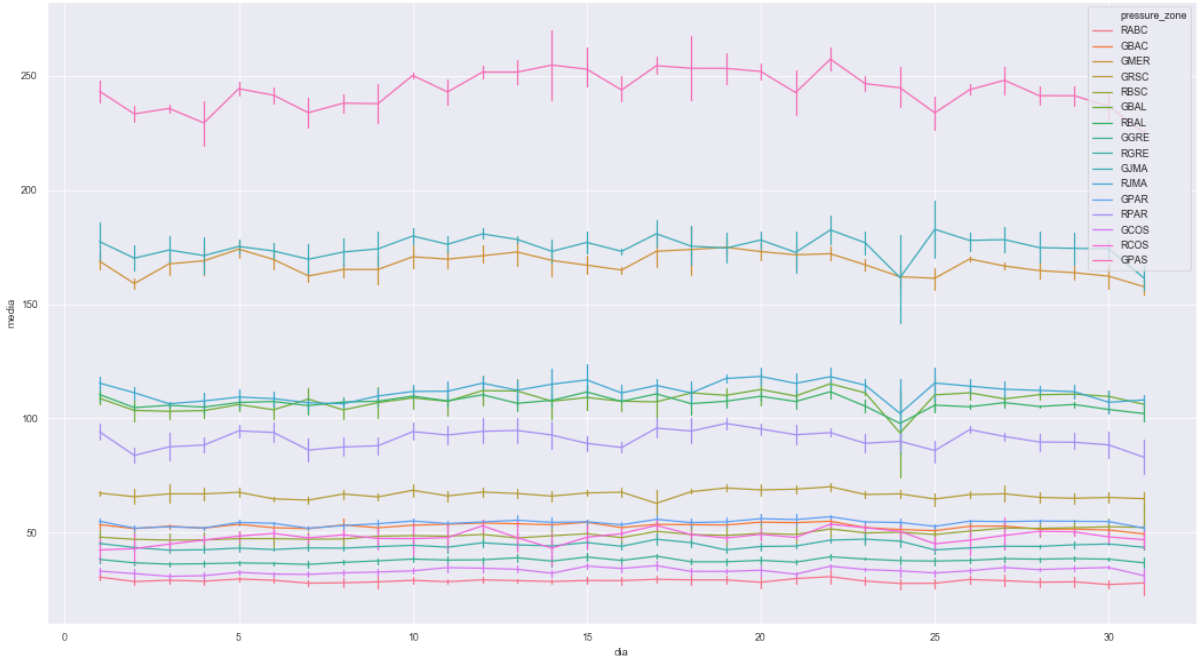
No Gráfico 4 e Gráfico 5 estão os valores mínima e média de todas as zonas de pressão. Com estas figuras é possível observar que os valores se mantêm em um padrão ao longo do mês, apenas nos valores de mínima que houve uma variação maior em alguns dias, o que possivelmente são anomalias na rede durante a noite. As barras apresentadas nesses gráficos são referentes ao intervalo de confiança da média calculada para cada ponto.

Gráfico 3 – Histograma da Zona de Pressão GMER.



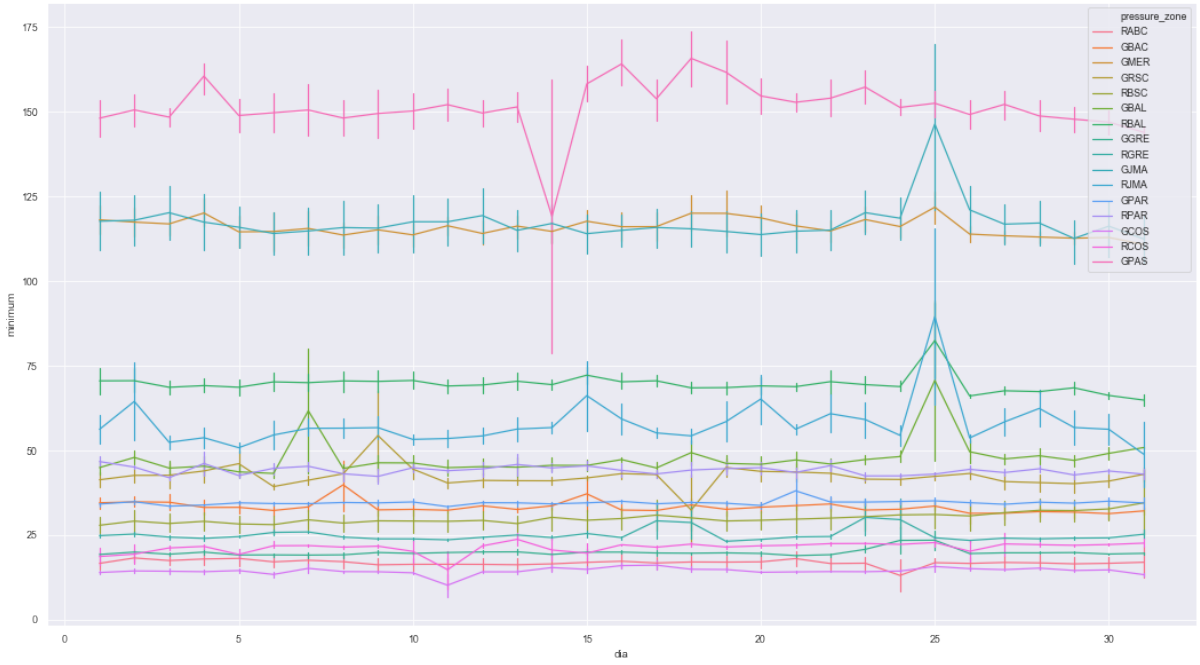
Fonte: Autoria própria.

Gráfico 4 – Média das dezesseis zonas de pressão.



Fonte: Autoria própria.

Gráfico 5 – Mínima das dezesseis zonas de pressão.



Fonte: Autoria própria.

4 IMPLEMENTAÇÃO

Para demonstrar a aplicabilidade das técnicas de AutoML a implementação da proposta é composta por três fases: (i) Implementação dos algoritmos sem usar otimização de parâmetros; (ii) Implementação da aprendizagem de máquina automatizada (AutoML) utilizando a biblioteca Auto-Sklearn; (iii) Otimização automática de hiper-parâmetros das técnicas SOM e LOF.

Nestas três etapas primeiramente aplicamos de forma simples as bibliotecas de aprendizagem de máquina SOM e LOF, apresentamos uma comparação destas com um algoritmo especialista e uma técnica estatística. Então na próxima etapa implementamos o AutoML e comparamos com as técnicas da etapa anterior. Por fim aplicamos o SOM e LOF utilizando uma abordagem semi-supervisionada e otimizamos os seus hiper-parâmetros por zona de pressão utilizando *Grid Search* (YU; ZHU, 2020).

4.1 SEM OTIMINIZAÇÃO

A primeira fase da implementação da proposta considera uma abordagem de implementação simples, sem recursos de otimização automática das tarefas. Esta implementação se baseia em aplicar algoritmos de detecção de outliers sobre os dados, para nos familiarizarmos com os dados e os problemas, desenvolvemos implementações iniciais da arquitetura proposta usando as técnicas SOM, LOF, Z-Score e o algoritmo Especialista, descritos nas próximas subseções.

A partir da base de dados relacional reproduzimos os gráficos usados atualmente pelos especialistas. O gráfico gerado através da base de dados relacional pôde ser gerado com histórico de três meses atrás, como pode ser visto no Gráfico 6 (atualmente os especialistas usam apenas um mês). A sombra gerada no gráfico é referente ao intervalo de confiança da média calculada para cada ponto.

Nesta fase utilizamos duas estratégias para avaliar o resultado dos algoritmos: uma manual, comparando gráficos similares aos usados atualmente pelos especialistas; e outra automática, usando os dados de operações de manutenção como indicadores (imprecisos) de falhas para cálculo de Precisão, Recall e F-Score.

A precisão é a medida da proporção de acertos dentro do conjunto retornado pelas técnicas. O Recall é a proporção de acertos dentro do conjunto total de itens corretos. Normalmente, conforme o Recall aumenta, a precisão diminui. F-Score é uma métrica que combina a

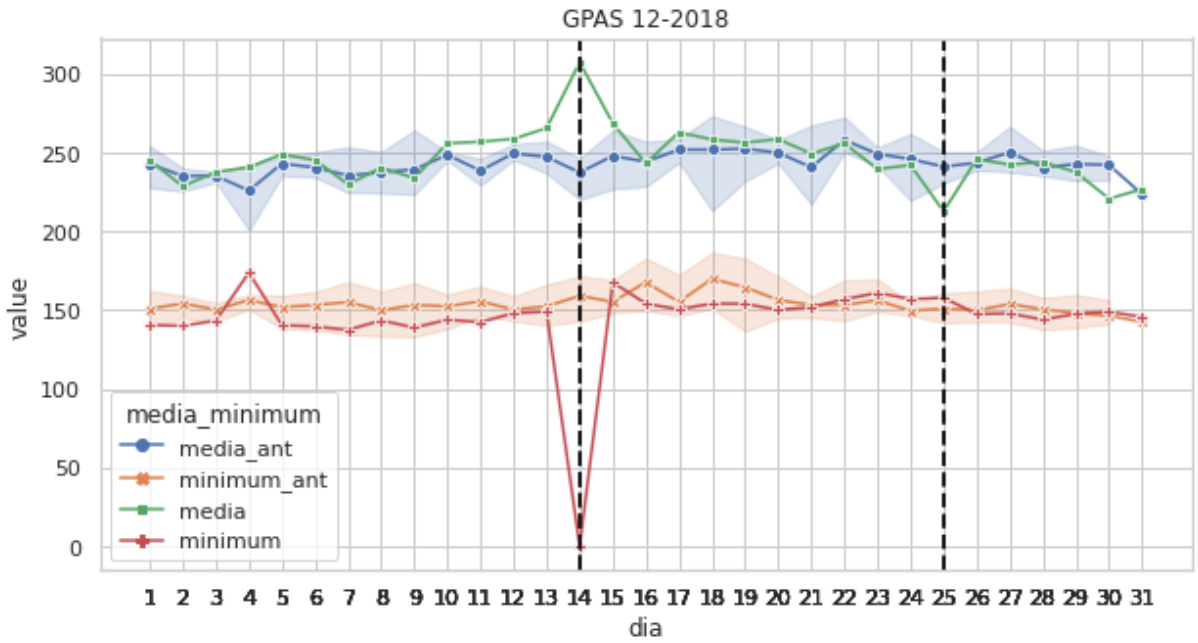
Precisão e o Recall, fazendo um balanceamento entre as duas através de um fator (CHINCHOR, 1992). Neste trabalho utilizamos a média harmônica como fator para obtermos um resultado equilibrado.

4.1.1 Algoritmos de Detecção Automática

A preparação inicial para aplicação das técnicas é um mapeamento dos dados por zonas de pressão com suas devidas entradas e saídas. Para tal, é realizada uma busca na base de dados selecionando os valores agrupados por dia do mês, efetuando uma média diária do fluxo e do fluxo mínimo noturno por zona de pressão. Obteve-se também o dia da semana como *feature* de contextualização. A princípio aplicamos as técnicas Z-Score, SOM e LOF para detecção automática de outliers sobre os dados das 16 zonas de pressão selecionadas.

Os resultados (Gráficos 6, 7 e 8) são exibidos em uma janela mensal com os outliers marcados com uma reta pontilhada na vertical. Este formato torna possível a comparação dos valores pelo agrupamento por dia do mês. As técnicas SOM e LOF também utilizam a variável dia da semana, além do dia do mês, no entanto não conseguimos visualizar a comparação da dimensão dia da semana nesse formato de exibição.

Gráfico 6 – Z-Score aplicado na mínima e média com pontuação acima de 2 (linhas tracejadas representam outliers detectados).

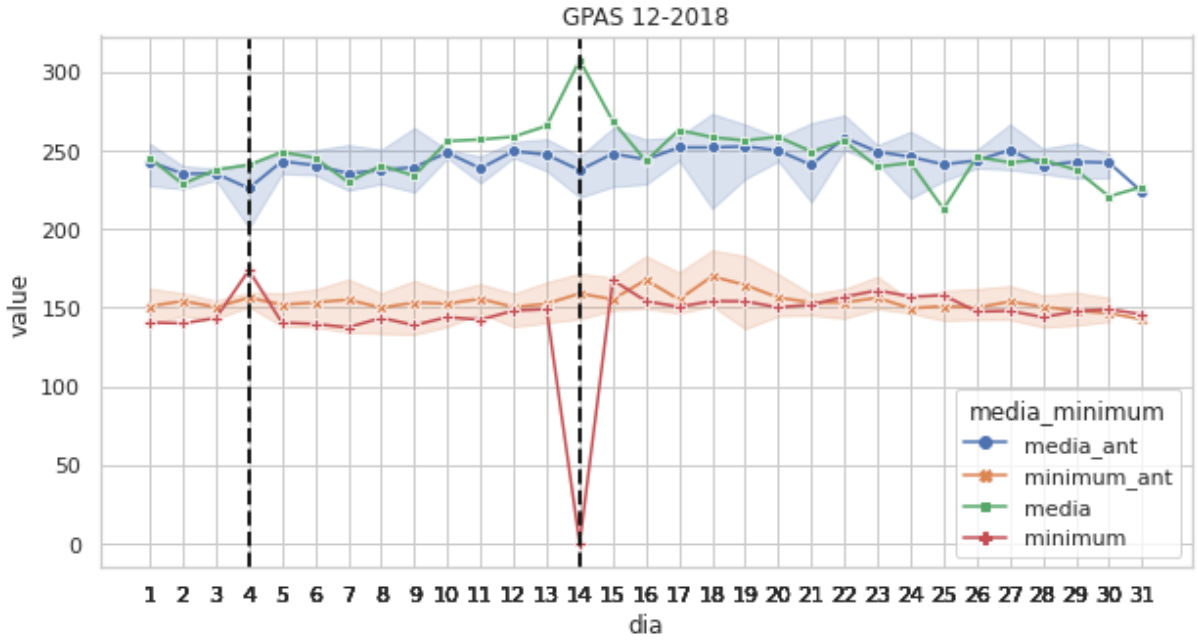


Fonte: Autoria própria.

A pontuação de Z-Score de cada valor é calculada subtraindo do valor diário o valor da

média geral. Esta diferença é dividida pelo desvio padrão. A aplicação do Z-Score é composta por duas etapas: primeiramente é aplicado o cálculo da pontuação no valor do fluxo mínimo noturno, considerando outliers os valores com pontuação acima de 2. Da mesma forma é efetuado no valor de fluxo médio diário. Então foram mesclados os resultados das duas etapas e apresentados no Gráfico 6. Ou seja, esta técnica utilizou apenas duas variáveis, pois não levou em consideração o dia da semana ou dia do mês.

Gráfico 7 – LOF aplicado com 10 vizinhos na zona de pressão GPAS, contaminação 0,05 (linhas tracejadas representam outliers detectados).

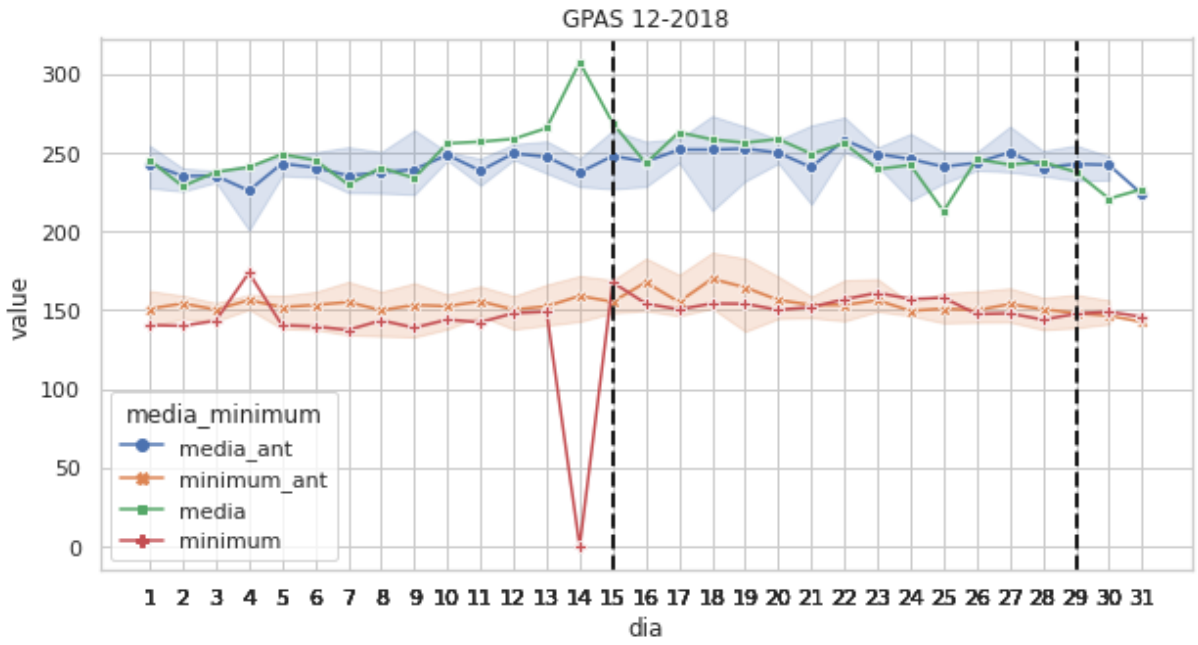


Fonte: Autoria própria.

Para as técnicas SOM e LOF são utilizadas as variáveis média diária, fluxo mínimo noturno, dia da semana e dia do mês. É importante utilizar tanto o dia do mês quanto do dia da semana para capturar informação de contexto sazonais de cada zona de pressão. Na aplicação do LOF são utilizados 10 vizinhos e uma contaminação de 0,05 (Gráfico 7).O SOM é aplicado com a largura do mapa igual a 15, uma contaminação de 0,05 e os resultados exibidos no Gráfico 8.

Nos Gráficos 6, 7 e 8 estão os valores de quatro meses apresentados em uma janela de um mês. Conforme exibe a legenda, os valores em verde e vermelho são referentes ao mês de referência (12-2018), os valores em laranja e azul são dos três meses anteriores ao mês de referência. As linhas pontilhadas na vertical são os dias do mês atual identificados como outlier pelas técnicas aplicadas. Nestes gráficos podemos observar dois picos que deveriam ser detectados como outliers, nos dias 4 e 14. Há outros pontos que podem ser identificados com outliers, como 25 e 30, no entanto, esses dias estão com os valores variando para baixo, isso não

Gráfico 8 – SOM aplicado com tamanho de mapa 15 na mínima e média (linhas tracejadas representam outliers detectados).



Fonte: Autoria própria.

deveria ser considerado outlier no nosso problema específico, pois a redução da vazão é desejada. As técnicas Z-Score e LOF detectaram corretamente o dia 14 como outlier e apenas o LOF detectou o dia 4. Os falsos positivos detectados poderiam ser minimizados usando estratégias como engenharia de *features*. Este porém não é o foco desta proposta.

As comparações feitas manualmente usando os gráficos demonstram que as técnicas automáticas têm potencial para identificação de situações anômalas. Porém, para uma análise mais completa, implementamos um algoritmo que simula os critérios usados atualmente pelos especialistas. Este algoritmo é apresentado na próxima subseção.

4.1.2 Algoritmo Especialista

Para simular a estratégia atual de detecção de vazamentos em Curitiba, apresentamos nesta seção um algoritmo que replica a atuação do especialista analisando o gráfico. O objetivo é ter uma referência para a comparação das técnicas. O algoritmo é definido a seguir.

Primeiro é definida a fórmula da função p (proporção), que retorna a proporção entre os valores lidos quando o valor de x é maior ou igual ao valor de y. Na comparação entre as duas variáveis de fluxo de água, somente é utilizado o valor de p quando há um aumento de vazão,

pois a redução de vazão é desejável e não configura um vazamento.

$$p(x,y) = \begin{cases} 0, & \text{if } x < y \\ (\frac{x}{y}), & \text{if } x \geq y \end{cases} \tag{2}$$

Na sequência é definido a fórmula do *score* do especialista (Specialist) pela função s. Onde α é a média dos valores de fluxo de água, β é a mínima, *i* é o índice do dia em análise, *i-1* é o índice do dia anterior e *i-30* é o índice do mês anterior. Os vetores $\overrightarrow{\alpha}$ e $\overrightarrow{\beta}$ representam as séries temporais das leituras.

$$s(\overrightarrow{\alpha}, \overrightarrow{\beta}, i) = p(\alpha_i, \alpha_{i-1}) + p(\alpha_i, \alpha_{i-30}) + p(\beta_i, \beta_{i-1}) + p(\beta_i, \beta_{i-30}), \text{for } i = 0..n \tag{3}$$

O algoritmo especialista (Specialist) é utilizado no mesmo padrão das técnicas LOF, Z-Score e SOM, com as mesmas variáveis, zonas de pressão, dados do mês de referência e os três meses anteriores. O *score* do algoritmo especialista é a soma das proporções, apenas quando é positiva, das variáveis mínima e média em relação ao dia anterior e ao mesmo dia do mês anterior. O algoritmo não leva em consideração o dia da semana.

4.1.3 Análise dos Resultados

Nesta subseção apresentamos resultados de qualidade dos algoritmos em termos de Precisão, Recall, F-Score e outras. Os eventos de outlier são representados pelo conjunto de dados de manutenção. Como estes dados de validação são imprecisos, é de se esperar que os valores de qualidade sejam baixos. Porém, o foco neste trabalho é na comparação entre as técnicas, logo os valores específicos alcançados pelas métricas são de menor importância. Ao aplicar os algoritmos, usamos diversas combinações de hiper-parâmetros.

A fim de padronizar a execução das técnicas definimos dois hiper-parâmetros, tamanho e contaminação, comum entre todas as técnicas descritas nas duas subseções anteriores (LOF, Z-Score, SOM e Especialista). Para o LOF, o parâmetro de tamanho é a quantidade de vizinhos, para o SOM, o tamanho do mapa e, para o Z-Score e o Specialist, o valor limite.

Na preparação dos dados para execução comum entre as técnicas montamos dois vetores para cada zona de pressão: o vetor de predição (contendo a saída dos algoritmos) e o vetor de validação (contendo dados de manutenção). O período usado foi de setembro a dezembro de 2018.

As métricas aplicadas nesse trabalho são a Correlação, Acurácia, Precisão, Recall, Support e F-Score. A ordenação e comparação entre as técnicas são utilizando o F-Score.

Com um padrão de métricas definido, com os vetores de dados preparados e as técnicas escolhidas, os seguintes valores para os hiper-parâmetros foram testados: SOM 10, 20, 50, 100 largura dos mapas; LOF 2, 3, 4, 5, 6, 7, 8, 9, 10 vizinhos; Z-Score com limite de 1.5, 2 e 2.5 sigmas.

Além dos exibidos nesta seção ainda foram testados outros hiper-parâmetros, como variação da função de vizinhança e a taxa de aprendizado do SOM, número de vizinhos do LOF, além de outras combinações entre as variáveis média, fluxo mínimo noturno, dia da semana e dia do mês.

Tabela 2 – Técnicas com maior F-Score para cada zona de pressão.

Z.P	Corr	Acurracy	Precision	Recall	Fscore	Support	Técnica	Cont
GBAC	0,21	0,85	0,43	0,18	0,25	17	SOM	0,05
GBAL	0,22	0,80	0,26	0,43	0,32	14	SPECIALIST	0,5
GCOS	0,03	0,72	0,24	0,16	0,19	25	SPECIALIST	0,5
GGRE	0,08	0,86	0,18	0,14	0,15	22	Z-SCORE	1,5
GJMA	0,12	0,76	0,43	0,11	0,17	28	LOF	0,05
GMER	0,01	0,58	0,40	0,12	0,19	96	Z-SCORE	1,5
GPAR	0,09	0,82	0,29	0,11	0,15	19	LOF	0,05
GPAS	0,12	0,66	0,57	0,10	0,16	42	LOF	0,05
GRSC	0,00	0,94	0,00	0,00	0,00	0	SOM	0,05
RABC	0,13	0,86	0,29	0,14	0,19	14	SOM	0,05
RBAL	0,19	0,84	0,43	0,16	0,23	19	LOF	0,05
RBSC	0,13	0,83	0,19	0,27	0,22	22	Z-SCORE	1,5
RCOS	0,06	0,77	0,24	0,14	0,18	42	Z-SCORE	1,5
RGRE	0,34	0,80	1,00	0,14	0,25	28	SPECIALIST	0,6
RJMA	0,10	0,75	0,43	0,10	0,16	30	LOF	0,05
RPAR	0,06	0,70	0,32	0,16	0,22	62	Z-SCORE	1,5

Fonte: Autoria própria.

Tabela 3 – Tabela de comparação dos valores de F-Score entre as técnicas

Técnica	Máximo	Mínimo	Média
SOM	0,25	0,08	0,15
LOF	0,29	0,07	0,15
Z-SCORE	0,23	0,06	0,14
SPECIALIST	0,32	0,07	0,15

Fonte: Autoria própria.

As tabelas 2 e 3 apresentam os resultados das execuções para as variações de hiper-parâmetros citadas anteriormente. Podemos observar que os valores de qualidade apresentados são baixos, os dados imprecisos de manutenção, por não representarem totalmente um outlier, são o principal motivo dessa baixa qualidade. Contudo, o foco desta proposta está na comparação

entre as técnicas, e devido todas utilizarem os mesmos dados para validação, a qualidade destes dados não influencia na escolha da melhor abordagem.

A tabela 2 apresenta o cálculo das métricas agrupados pelas dezesseis zonas de pressão. Para cada zona de pressão é exibida a técnica com maior F-Score. Como pode-se observar nesta tabela, não há dominância de nenhuma técnica. Na tabela 3 apresentamos os mesmos resultados agrupados por técnica, com seu respectivo maior F-Score, menor e a média. A máxima do algoritmo especialista ficou um pouco acima das outras, no entanto na média as técnicas ficaram empatadas. Apenas o Z-Score ficou 0,01 abaixo.

Na próxima seção apresentamos a aplicação do AutoML e comparamos com as técnicas descritas acima. O objetivo é determinar se existem vantagens práticas na adaptação do problema de detecção de outliers em uma técnica supervisionada e otimizada por AutoML.

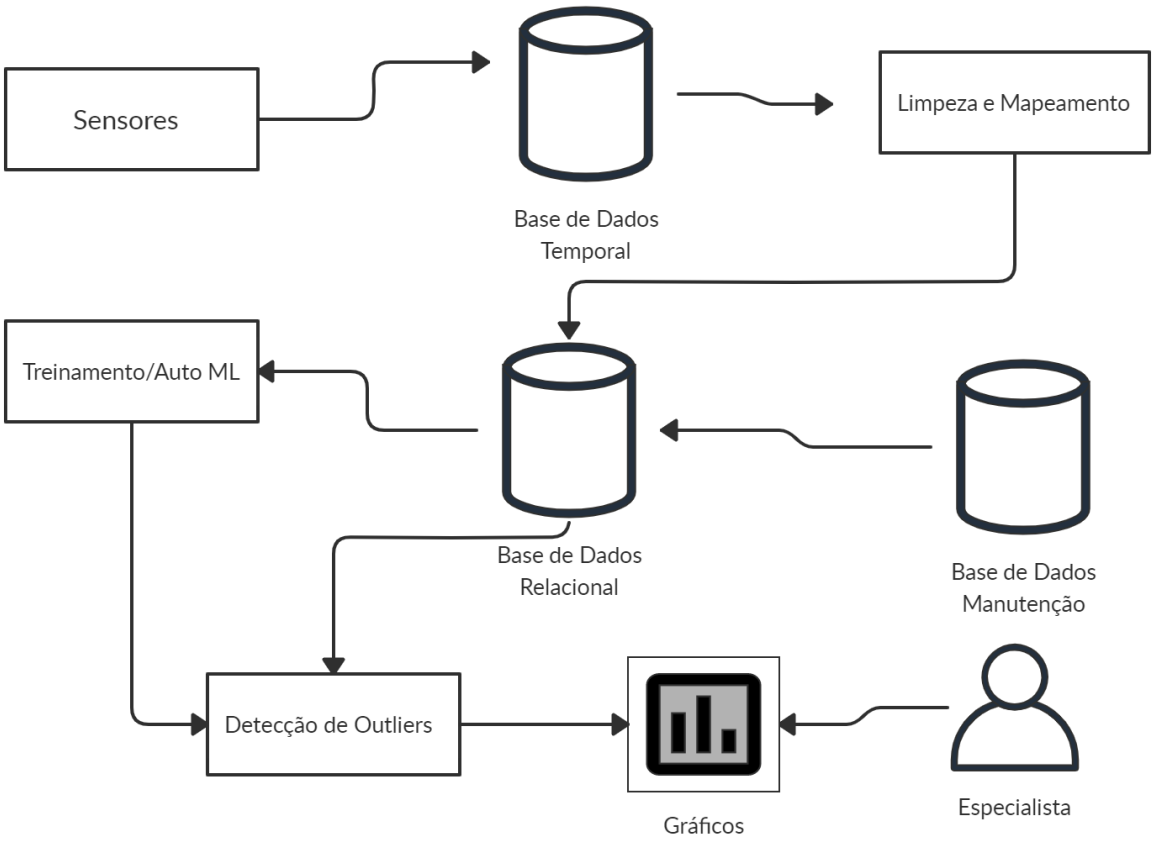
4.2 AUTOML USANDO AUTO-SKLEARN

Conforme descrito anteriormente, o objetivo desta etapa é apresentar uma solução viável de aplicação para detecção automática dos vazamentos ou eventos anômalos no fluxo de água de dezesseis zonas de pressão dentro do sistema de distribuição de água de Curitiba. Desta forma, esta seção apresenta uma primeira abordagem para a proposta central desta dissertação, de tratar o problema de detecção de outliers em sistemas de distribuição de água. A Figura 5 mostra a arquitetura desta solução.

Na Figura 5 temos o preenchimento da base de dados relacional através de outras duas bases de dados. A base de dados temporal para obtenção dos dados de fluxo e a base de dados de manutenção. A base de dados temporal é preenchida com dados de sensores de fluxo de água instalados na rede, com frequência em tempo real. Os dados de fluxo podem conter ruídos, e devido à falta de relacionamento entre as *tags* é necessária a etapa de limpeza e mapeamento. A base de dados de manutenção tem seu preenchimento através de um sistema em plataforma mainframe pelos profissionais responsáveis de cada zona de pressão. Para treinamento e otimização do modelo no formato AutoML, os dados são selecionados da base relacional por zona de pressão no período de oito meses com agrupamento diário. As técnicas são aplicadas a novos dados da base de dados relacional e os resultados apresentados aos especialistas através de relatórios (gráficos).

As técnicas apresentadas na seção 4 utilizaram para treinamento todos os dados de fluxo de água do período selecionado. As técnicas apontavam os outliers com base no hiper-parâmetro

Figura 5 – Arquitetura para detecção de outliers.



Fonte: Autoria própria.

de contaminação (que especifica a porcentagem de dados considerados anômalos). A saída dos algoritmos (outliers detectados) eram usadas para a comparação com os dados de manutenção do mesmo período. Ou seja, para treinamento eram utilizados todos os dados e para validação utilizávamos os dados de manutenção. Contudo, para aplicação das técnicas de aprendizado supervisionado do AutoML realizamos uma divisão entre conjunto de treinamento (75%) e testes (25%) em ambos os conjuntos de dados (dados de fluxo e dados de manutenção).

Para a aplicação do AutoML, empregamos a biblioteca Auto-Sklearn (FEURER *et al.*, 2015) descrita na subseção 2.6.

Dado o conjunto de treinamento, realizamos a aplicação simples do AutoML para todas as zonas de pressão. Contudo, os resultados iniciais da aplicação do AutoML foram insatisfatórios. Com o objetivo de melhorar os resultados do AutoML, tomamos algumas medidas como: a variação na qualidade dos resultados devido à aleatoriedade na seleção dos dados de treinamento e testes, aumentar as informações de contexto e melhorar a distinção entre as classes.

O aumento do conjunto de dados de treinamento pode melhorar significativamente a qualidade do modelo. Portanto dobramos a quantidade de dados, de quatro para oito meses de tempo utilizado (de maio a dezembro de 2018).

Para reduzir a variação na qualidade dos dados selecionados para treinamento, implementamos o método K-Fold de validação cruzada. O K-Fold, através de várias execuções, garante que todos os dados sejam utilizados tanto para treinamento quanto para testes.

Utilizamos o K-Fold de forma a dividir o conjunto de dados em quatro partições (*folds*). Em cada execução eram utilizados 25% dos dados para o conjunto de testes, mutuamente exclusivos. A aplicação do Auto-Sklearn e cálculo das métricas passaram a usar divisão por K-Fold. Ao fim é utilizada a média aritmética de quatro *folds*.

Conforme demonstramos na seção 4.1 as técnicas tendem a detectar as quedas na vazão como outlier. A adição de novas *features* pode auxiliar na aquisição de informações de contexto pelas técnicas. Partindo da forma que o algoritmo especialista funciona, calculando a proporção entre dois dias e utilizando apenas as proporções positivas, adicionamos novas variáveis seguindo a mesma abordagem, no entanto calculando a diferença em vez da proporção.

As *features* adicionadas são referentes ao dia da semana e os valores de mínima e média. Utilizamos o valor da semana anterior para mínima e média, com a respectiva diferença. Da mesma forma calculamos os valores das duas semanas e três semanas anteriores. Isto gerou as seguintes variáveis: a) mínima da semana anterior; b) diferença para a mínima da semana anterior; c) média da semana anterior; d) diferença para a média da semana anterior; e) mínima de duas semanas atrás; f) diferença para a mínima de duas semanas atrás; e) média de duas semanas atrás; g) diferença para a média de duas semanas atrás; h) mínima de três semanas atrás; i) diferença para a mínima de três semanas atrás; j) média de três semanas atrás; k) diferença para a média de três semanas atrás. Na tabela 4 é apresentado quais *features* são utilizadas por cada modelo.

Executamos o Auto-Sklearn com as combinações de tamanho 1, 50, 100 e 200 para o parâmetro *ensemble size*. E com 30, 60, 90, 180, 300 segundos para o parâmetro de tempo de execução. Devido ao aumento da janela de tempo, o aumento do número de variáveis, o aumento do número de combinações com o K-Fold e hiper-parâmetros do Auto-Sklearn, reduzimos o tempo de execução, pois o padrão da ferramenta de 1 hora se tornou inviável. Com as execuções destas combinações foram montadas as tabelas 5 e 6.

Na tabela 5 apresentamos os valores agrupados por zonas de pressão e técnica com

Tabela 4 – Features utilizadas pelos modelos.

Features	Z-Score	Specialist	SOM/LOF (Seção 4.1)	AutoML	SOM/LOF (Seção 4.2 e 4.3)
Vazão mínima noturna.	x	x	x	x	x
Média da vazão diária.	x	x	x	x	x
Dia do mês.		x	x	x	x
Dia da semana.			x	x	x
Mínima da semana anterior.				x	x
Diferença para a mínima da semana anterior.				x	x
Média da semana anterior.				x	x
Diferença para a média da semana anterior.				x	x
Mínima de duas semanas atrás.				x	x
Diferença para a mínima de duas semanas atrás.				x	x
Média de duas semanas atrás.				x	x
Diferença para a média de duas semanas atrás.				x	x
Mínima de três semanas atrás.				x	x
Diferença para a mínima de três semanas atrás.				x	x
Média de três semanas atrás.				x	x
Diferença para a média de três semanas atrás.				x	x

Fonte: Autoria própria.

maior média de F-Score entre todas as execuções. Também é possível visualizar os parâmetros de tamanho e contaminação utilizados, além do maior e menor valor de F-Score que a técnica obteve para aquela zona de pressão. Com esses resultados podemos observar que o Auto-Sklearn ainda obteve um resultado inferior ao das outras técnicas, principalmente o SOM. Na maioria das zonas de pressão o SOM apresentou os melhores resultados.

Na tabela 6 apresentamos os valores agrupados por técnica. O maior e o menor valor para cada técnica, além da média de todas as combinações para todas as zonas de pressão. Para apresentar os resultados por técnica na tabela 6 foram removidos os valores de F-Score zerados na identificação dos mínimos. Os resultados do Auto-Sklearn continuaram apresentando alguns valores zerados. No momento da execução de alguns algoritmos o Auto-Sklearn exibia uma mensagem apontando que o resultado encontrado é inferior ao de um modelo randômico. Esta e outras questões com a ferramenta são discutidas na seção 5.

Nos resultados iniciais obtidos na seção 4.1 sem otimização de hiper-parâmetros as técnicas demonstraram a média de 0,15. Após as alterações realizadas, inclusão de mais variáveis

Tabela 5 – Técnicas com maior F-Score para cada zona de pressão.

Z.P	Téc.	Tam.	Cont.	Máx.	Mín.	Média
RABC	SPECIALIST	1	0,70	0,28	0,09	0,17
GBAC	SOM	10	0,05	0,31	0,18	0,24
GMER	Z-SCORE	1	2,00	0,67	0,45	0,54
RBSC	SPECIALIST	1	0,70	0,26	0,09	0,18
GBAL	SOM	10	0,05	0,24	0,13	0,19
RBAL	SOM	10	0,05	0,49	0,21	0,34
GGRE	SOM	10	0,05	0,20	0,15	0,18
RGRE	SOM	10	0,05	0,46	0,18	0,32
GJMA	SPECIALIST	1	0,50	0,48	0,23	0,38
RJMA	Z-SCORE	1	2,00	0,47	0,31	0,40
GPAR	SOM	10	0,05	0,31	0,21	0,27
RPAR	Z-SCORE	1	2,00	0,53	0,38	0,44
GCOS	SOM	10	0,05	0,45	0,20	0,32
RCOS	SOM	10	0,05	0,48	0,21	0,36
GPAS	SOM	10	0,05	0,58	0,44	0,52

Fonte: Autoria própria.

Tabela 6 – Tabela de comparação dos valores de F-Score entre as técnicas.

Téc.	Máximo	Mínino	Média
SOM	0,60	0,07	0,32
LOF	0,66	0,10	0,31
Z-SCORE	0,67	0,09	0,31
SPECIALIST	0,67	0,06	0,31
AUTOSKLEARN	0,53	0,07	0,26

Fonte: Autoria própria.

de contexto, implementação do K-Fold e aumento do conjunto de dados, os resultados dobraram os valores apresentados anteriormente.

Estes resultados indicam que os algoritmos SOM e LOF são mais capazes de capturar aspectos que determinam a anormalidade de uma observação. Uma provável razão para esta capacidade pode estar relacionada com a ênfase na distribuição espacial das densidades das observações, em contraste às subdivisões do espaço realizadas por algoritmos tradicionais de classificação. Para unir as vantagens dos algoritmos SOM e LOF com a eficiência do AutoML, desenvolvemos então um esquema semi-supervisionado para otimização de hiper-parâmetros, descrito na próxima seção.

4.3 OTIMIZAÇÃO DE HIPER-PARÂMETROS PARA O SOM E LOF NO ESTILO AUTOML

Em consequência dos resultados insatisfatórios obtidos com as técnicas de classificação no Auto-Sklearn, optamos por aplicar uma otimização de hiper-parâmetros sobre os algoritmos SOM e LOF. O objetivo é elaborar uma abordagem semi-supervisionada através da seleção dos melhores hiper-parâmetros utilizando os dados de manutenção. Nossa hipótese é que os

algoritmos tradicionais de detecção de outliers podem ser mais eficientes que os algoritmos de classificação usados no Auto-sklearn por darem mais ênfase à densidade das observações no espaço e não à busca de subdivisões do espaço em classes.

Primeiramente fizemos uma tentativa de integração do SOM e LOF com o Auto-Sklearn para realizar essa otimização automaticamente. Tais algoritmos não estão disponíveis na biblioteca Auto-Sklearn por serem não supervisionados. Esta integração não foi possível, devido a biblioteca não executar o método de predição das nossas técnicas (descrevemos com mais detalhes os problemas encontrados na seção 5). Portanto, a estratégia adotada se baseia na otimização manual dos hiper-parâmetros do SOM e do LOF usando um algoritmo de Grid Search.

Na execução padrão do SOM e LOF, os dados são passados em um único conjunto e através da contaminação são retornados os outliers desse conjunto. Desta forma o SOM e o LOF eram executados em todo o conjunto de dados, e para cada entrada obtinha-se uma saída (outlier ou inlier). Com isso os resultados referentes ao conjunto de testes eram comparados com os dados de manutenção. Contudo, com a experiência que obtivemos ao aplicar o Auto-sklearn observamos que essa não era uma abordagem igualitária, pois não é coerente realizar a comparação entre as técnicas que utilizam todos os dados para treinamento com outras que fazem a divisão entre conjunto de treinamento e testes.

Para sanar essa diferença nas abordagens foi realizada uma alteração na forma de execução do SOM e do LOF. O principal objetivo dessa alteração é que os dados de testes não sejam utilizados no treinamento para não enviesar as técnicas. Assim, a parte do processo referente à divisão dos conjuntos de dados é igual para ambas as abordagens (ou seja, os conjuntos de dados de vazão e de manutenção foram divididos em conjunto de treinamento e de testes nos mesmos moldes do K-Fold explicado na seção 4.2).

Para a adaptação dos algoritmos SOM e LOF à otimização automática de hiper-parâmetros, utilizamos duas estratégias diferentes: para o LOF foi utilizada uma função de detecção de novidades da própria biblioteca; para o SOM criamos uma função com base no *threshold* de treinamento.

O treinamento do LOF com a detecção de novidades ativada recebe os dados de treinamento do conjunto de vazão, não utilizando o conjunto de treinamento de manutenção (ou seja, um treinamento não supervisionado). Após realizado o treinamento é chamado o método de predição utilizando os dados de teste do conjunto de vazão. Com os resultados realizamos a

comparação com o conjunto de testes dos dados de manutenção (para calcular o F-Score usado como base para a otimização).

O processo desenvolvido manualmente para o SOM se inicia com a aplicação da técnica para a obtenção do fator de outlier (*outlier score*) para cada entrada do conjunto de treinamento. Com base na contaminação, calcula-se o *threshold* do fator de outlier para classificar as entradas do conjunto de testes. A saída do conjunto de testes é utilizada para calcular o F-Score junto ao conjunto de validação (dados de manutenção).

Para otimização dos hiper-parâmetros, implementamos um procedimento de Grid Search. Os hiper-parâmetros usados são descritos na tabela 7.

Tabela 7 – Valores utilizados para realizar o Grid Search nos parâmetros das técnicas SOM e LOF.

Técnica	Parâmetro	Valores	Iteração
SOM	Tamanho do mapa	44 a 66	2
SOM	Contaminação	0,06 a 0,2	0,02
SOM	Sigma	1 a 3	1
SOM	Função de Vizinhança	gaussian, bubble e triangle.	-
SOM	Taxa de Aprendizagem	3 a 6	1
LOF	Algoritmo	Auto	-
LOF	Tamanho da Folha	20	1
LOF	Quantidade de vizinhos	12 a 50	2
LOF	Contaminação	0,05 a 0,2	0,01
LOF	Métrica	cityblock, cosine, l1, manhattan, canberra, dice, jaccard, rogerstanimoto, russellrao, sokalmichener, sokalsneath.	-

Fonte: Autoria própria.

Os resultados estão apresentados nas tabelas 8 e 9. Param1 é referente ao sigma do SOM e Algoritmo do LOF. Param2 é referente a função de vizinhança do SOM e tamanho da folha do LOF. Param3 é referente a taxa de aprendizagem do SOM e a métrica do LOF. O parâmetro tamanho da folha do LOF é sempre 20 e o algoritmo do LOF é sempre automático.

Na tabela 9 a média, mínima e máxima foram calculadas referente apenas às execuções vencedoras de cada zona de pressão, devido à grande variação no resultado ocasionada por numerosas combinações de hiper-parâmetros durante a otimização.

A alteração na abordagem foi positiva, pois além de igualar a comparação com os métodos de classificação, também habilitou a possibilidade prática de aplicação da detecção de outliers em novos dados, podendo ser utilizada em tempo real.

Não utilizar todo o conjunto de dados para treinamento poderia reduzir a qualidade do modelo, fazendo com que os resultados fossem inferiores, no entanto os resultados após essa alteração e a otimização dos hiper-parâmetros foram positivos (Tabela 10).

Tabela 8 – Técnicas com maior média de F-Score para cada zona de pressão.

Z.P	Téc.	Tam.	Cont.	param1	param2	param3	Máx	Mín.	Média
RABC	SOM	52	0,06	1	gaussian	4	0,52	0,10	0,22
GBAC	SOM	54	0,18	1	gaussian	3	0,67	0,26	0,39
GMER	LOF	42	0,12	auto	20	russellrao	0,69	0,44	0,54
RBSC	LOF	14	0,2	auto	20	canberra	0,30	0,15	0,23
GBAL	SOM	44	0,2	1	bubble	5	0,33	0,13	0,20
RBAL	SOM	54	0,1	1	bubble	5	0,51	0,21	0,37
GGRE	SOM	56	0,18	1	bubble	3	0,28	0,18	0,23
RGRE	LOF	28	0,2	auto	20	rogerstanimoto	0,50	0,19	0,36
GJMA	LOF	46	0,13	auto	20	russellrao	0,52	0,24	0,43
RJMA	LOF	44	0,2	auto	20	russellrao	0,51	0,32	0,43
GPAR	LOF	24	0,18	auto	20	rogerstanimoto	0,44	0,23	0,33
RPAR	SOM	46	0,16	1	bubble	6	0,53	0,40	0,47
GCOS	LOF	28	0,19	auto	20	dice	0,43	0,30	0,37
RCOS	SOM	50	0,18	1	gaussian	5	0,48	0,32	0,41
GPAS	LOF	20	0,13	auto	20	russellrao	0,58	0,44	0,52

Fonte: Autoria própria.

Tabela 9 – Tabela de comparação dos valores de F-Score entre as técnicas.

Téc.	Máximo	Mínimo	Média
SOM	0,67	0,10	0,36
LOF	0,69	0,10	0,35

Fonte: Autoria própria.

Tabela 10 – Tabela de comparação da média de F-Score entre as seções.

Téc.	Seção 4.1	Seção 4.2	Seção 4.3
SOM	0,15	0,32	0,36
LOF	0,15	0,31	0,35

Fonte: Autoria própria.

5 DISCUSSÃO

Os principais desafios encontrados na implementação derivam de duas características centrais do contexto desta proposta: (i) dados de validação imprecisos e (ii) dependência do contexto na determinação dos outliers.

Os dados de validação imprecisos (registros de manutenção parcialmente relacionados com vazamentos) geraram problemas na distinção entre as classes (outlier ou não outlier) e um baixo valor de F-Score (entre 0,3 e 0,7). A linha tênue de distinção das classes fez com que, em alguns casos, os algoritmos de classificação do Auto-Sklearn não retornassem nenhum resultado como outlier, e como consequência a apresentação de um F-Score zerado.

A justificativa para utilizar os dados imprecisos de manutenção é a aplicabilidade em um cenário real, visto que os dados foram extraídos diretamente do sistema em produção.

A segunda questão, a dependência de contexto na determinação dos outliers, diz respeito à importância de se analisar características de observações similares. Nos algoritmos SOM e LOF, o contexto é capturado no espaço multidimensional formado pelas variáveis. Nestes algoritmos, o ponto em análise é caracterizado como outlier dependendo da densidade relativa das observações vizinhas. Este tipo de análise espacial não é facilmente realizada com algoritmos de classificação comumente usados em AutoML. Para resolver esta questão adicionamos novas variáveis (*features*) de contexto e tentamos realizar a integração das técnicas SOM e LOF no Auto-Sklearn.

A geração automática de novas *features* propunha realizar a mescla entre as já existentes através de operações matemáticas simples, no entanto a partir da experiência com o algoritmo especialista criamos essas *features* manualmente calculando a diferença para um valor anterior da mesma, no caso foram adicionadas a diferença de 7 dias atrás, 14 dias atrás e 21 dias atrás. Com isso, as técnicas puderam obter os padrões sazonais dentro desta janela de tempo.

A integração dos algoritmos SOM e LOF com a ferramenta Auto-Sklearn traria benefícios como obter os melhores hiper-parâmetros para estas técnicas, além de resolver o problema do Auto-Sklearn não apresentar resultados no conjunto de outliers.

Para a integração é necessária a implementação de alguns métodos conforme está presente na documentação do Auto-Sklearn. Porém, apesar de cumprirmos todos os requisitos tivemos problemas com o método de predição implementado, este nunca era chamado durante a execução e desta forma não apresentava os resultados esperados, mantendo o mesmo padrão

anterior de não apresentar resultados para a classe de outlier.

6 CONCLUSÃO

O problema das perdas de água é importante e deve ser levado a sério pelas empresas de saneamento. Para isso, o monitoramento diário com detecção de outliers pode ser fundamental para a redução desse índice.

O objetivo prático é a redução do tempo de detecção de vazamentos e a detecção de pequenos vazamentos que podem existir na rede e não foram detectados ainda. Propomos para resolver esse problema a aplicação de detecção de outlier com otimização automática de hiper-parâmetros com aprendizagem semi-supervisionada.

Demonstramos que as técnicas não supervisionadas SOM e LOF aplicadas de forma semi-supervisionada obtiveram melhores resultados em comparação ao modelo criado pelo AutoML através da ferramenta Auto-Sklearn.

Este trabalho pode ser aplicado de forma prática e tem potencial de causar um impacto positivo na redução das perdas de água de Curitiba e Região Metropolitana. A proposta também é adaptável para outros contextos, sejam em cidades diferentes ou problemas similares (e.g. distribuição de energia, infraestruturas de comunicação, etc.).

Como trabalhos futuros pretendemos: (i) Incluir mais features, por exemplo: datas que são feriados, dados de precipitação de chuva, temperatura, etc; (ii) Aplicar engenharia de features automatizada; (iii) Implementar técnicas mais avançadas de busca no espaço de hiper-parâmetros; (iv) Implementar aprendizagem adaptativa com feedback de especialistas; (v) Incluir os dados do sistema web (citado na subseção 3.1.3) na etapa de treinamento das técnicas; (vi) Diminuir a granularidade dos dados de fluxo, atualmente estamos utilizando valores diários, grande parte das leituras são gravadas em tempo real, isso possibilita a utilização de valores horários; (vii) Aumentar a quantidade de zonas de pressão; (viii) Ajustar o alfa do F-Score para tratar os falsos positivos.

REFERÊNCIAS

AKSELA, K; AKSELA, M; VAHALA, R. Leakage detection in a real distribution network using a som. **Urban Water Journal**, Taylor & Francis, v. 6, n. 4, p. 279–289, 2009.

BENGHI, Felipe Marx. **Visual Analytics e Outlying Aspect Mining: Contextualização de Anomalias Considerando Questões Temporais e Multidimensionais**. 2020. Dissertação (Mestrado) — Universidade Tecnológica Federal do Paraná, 2020.

CHAN, TK; CHIN, Cheng Siong; ZHONG, Xionghu. Review of current technologies and proposed intelligent methodologies for water distributed network leakage detection. **IEEE Access**, IEEE, v. 6, p. 78846–78867, 2018.

CHANDOLA, Varun; BANERJEE, Arindam; KUMAR, Vipin. Anomaly detection: A survey. **ACM computing surveys (CSUR)**, ACM, v. 41, n. 3, p. 15, 2009.

CHINCHOR, Nancy. Muc-4 evaluation metrics. *In*: **Fourth Message Uunderstanding Conference (MUC-4): Proceedings of a Conference Held in McLean, Virginia, June 16-18, 1992**. *[S.l.: s.n.]*, 1992.

FEURER, Matthias; KLEIN, Aaron; EGGENSPERGER, Katharina; SPRINGENBERG, Jost; BLUM, Manuel; HUTTER, Frank. Efficient and robust automated machine learning. *In*: CORTES, C.; LAWRENCE, N. D.; LEE, D. D.; SUGIYAMA, M.; GARNETT, R. (Ed.). **Advances in Neural Information Processing Systems 28**. Curran Associates, Inc., 2015. p. 2962–2970. Disponível em: http://papers.nips.cc/paper/5872-efficient-and-robust-automated-machine-learning.pdf.

GUPTA, Manish; GAO, Jing; AGGARWAL, Charu; HAN, Jiawei. Outlier detection for temporal data. **Synthesis Lectures on Data Mining and Knowledge Discovery**, Morgan & Claypool Publishers, v. 5, n. 1, p. 1–129, 2014.

HUTTER, Frank; KOTTHOFF, Lars; VANSCHOREN, Joaquin. **Automated Machine Learning**. *[S.l.]*: Springer, 2019.

JUNIOR, Orlando Stein. **Detecção de Variação de Perfil de Consumo de Energia Elétrica para o Grupo A, Utilizando Mapas Auto-Organizáveis e Mineração de Dados**. 2018. Dissertação (Mestrado) — Instituto de Tecnologia para o Desenvolvimento, 2018.

KATZ, Gilad; SHIN, Eui Chul Richard; SONG, Dawn. Explorekit: Automatic feature generation and selection. *In*: IEEE. **2016 IEEE 16th International Conference on Data Mining (ICDM)**. *[S.l.]*, 2016. p. 979–984.

KAUL, Ambika; MAHESHWARY, Saket; PUDI, Vikram. Autolearn—automated feature generation and selection. *In*: IEEE. **2017 IEEE International Conference on data mining (ICDM)**. *[S.l.]*, 2017. p. 217–226.

KOHONEN, Teuvo; HONKELA, Timo. Kohonen network. **Scholarpedia**, v. 2, n. 1, p. 1568, 2007.

KREYSZIG, Erwin. Advanced engineering mathematics 10th edition. Publisher John Wiley & Sons, 2009.

LAMBERT, Allan; HIRNER, Wolfram. **Losses from Water Supply Systems: Astandard Terminology and Recommended Performance Measures**. *[S.l.]*: IWA, 2000.

LI, Rui; HUANG, Haidong; XIN, Kunlun; TAO, Tao. A review of methods for burst/leakage detection and location in water distribution systems. **Water Science and Technology: Water Supply**, IWA Publishing, v. 15, n. 3, p. 429–441, 2015.

LOUREIRO, Dália; AMADO, Conceição; MARTINS, André; VITORINO, Diogo; MAMADE, Aisha; COELHO, Sérgio Teixeira. Water distribution systems flow monitoring and anomalous event detection: A practical approach. **Urban Water Journal**, Taylor & Francis, v. 13, n. 3, p. 242–252, 2016.

MOUNCE, Stephen R; KHAN, Asar; WOOD, Alastair S; DAY, Andrew J; WIDDOP, Peter D; MACHELL, John. Sensor-fusion of hydraulic data for burst detection and location in a treated water distribution system. **Information Fusion**, Elsevier, v. 4, n. 3, p. 217–229, 2003.

PALAU, CV; ARREGUI, FJ; CARLOS, M. Burst detection in water networks using principal component analysis. **Journal of Water Resources Planning and Management**, American Society of Civil Engineers, v. 138, n. 1, p. 47–54, 2011.

PUUST, R; KAPELAN, Z; SAVIC, DA; KOPPEL, T. A review of methods for leakage management in pipe networks. **Urban Water Journal**, Taylor & Francis, v. 7, n. 1, p. 25–45, 2010.

ROMANO, M; KAPELAN, Z; SAVIĆ, DA. Real-time leak detection in water distribution systems. *In*: **Water Distribution Systems Analysis 2010**. *[S.l.]*: 12th Annual Conference on Water Distribution Systems Analysis (WDSA), 2010. p. 1074–1082.

ROMANO, Michele; KAPELAN, Zoran; SAVIĆ, Dragan A. Automated detection of pipe bursts and other events in water distribution systems. **Journal of Water Resources Planning and Management**, American Society of Civil Engineers, v. 140, n. 4, p. 457–467, 2012.

RUSSELL, Stuart J; NORVIG, Peter. **Artificial intelligence: a modern approach**. *[S.l.]*: Malaysia; Pearson Education Limited,, 2016.

THORNTON, Chris; HUTTER, Frank; HOOS, Holger H; LEYTON-BROWN, Kevin. Auto-weka: Combined selection and hyperparameter optimization of classification algorithms. *In*: **Proceedings of the 19th ACM SIGKDD international conference on Knowledge discovery and data mining**. *[S.l.: s.n.]*, 2013. p. 847–855.

TRATA BRASIL. Instituto trata brasil. **Perdas de água (SNIS 2017): Desafios para Disponibilidade Hídrica e Avanço da Eficiência do Saneamento Básico (2019)**, 2017.

TUGGENER, Lukas; AMIRIAN, Mohammadreza; ROMBACH, Katharina; LÖRWALD, Stefan; VARLET, Anastasia; WESTERMANN, Christian; STADELMANN, Thilo. Automated machine learning in practice: state of the art and recent results. *In*: IEEE. **2019 6th Swiss Conference on Data Science (SDS)**. *[S.l.]*, 2019. p. 31–36.

WU, Yipeng; LIU, Shuming. A review of data-driven approaches for burst detection in water distribution systems. **Urban Water Journal**, Taylor & Francis, v. 14, n. 9, p. 972–983, 2017.

YU, Tong; ZHU, Hong. Hyper-parameter optimization: A review of algorithms and applications. **arXiv preprint arXiv:2003.05689**, 2020.