Universidade Estadual de Campinas

Faculdade de Engenharia Química

Rodolfo Pinheiro da Cruz

Método Acústico e Algoritmos de Aprendizado de Máquinas Aplicados a Detecção e Localização de Vazamentos em Tubulações de Gás de Baixa Pressão

Campinas 2019

Rodolfo Pinheiro da Cruz

Método Acústico e Algoritmos de Aprendizado de Máquinas Aplicados a Detecção e Localização de Vazamentos em Tubulações de Gás de Baixa Pressão

Dissertação de mestrado apresentado à Faculdade de Engenharia Química da Universidade Estadual de Campinas como parte dos requisitos para obtenção do título de Mestre em Engenharia Química.

Orientadora: Profª Drª. Ana Maria Frattini Fileti

Este exemplar corresponde à versão final da dissertação defendida pelo aluno Rodolfo Pinheiro da Cruz e orientada pela profª. Drª. Ana Maria Frattini Fileti.

Campinas

2019

Ficha catalográfica
Universidade Estadual de Campinas
Biblioteca da Área de Engenharia e Arquitetura
Luciana Pietrosanto Milla - CRB 8/8129

C889m

Cruz, Rodolfo Pinheiro da, 1987-
Método acústico e algoritmos de aprendizado de máquinas aplicados a detecção e localização de vazamentos em tubulações de gás de baixa pressão / Rodolfo Pinheiro da Cruz. – Campinas, SP : [s.n.], 2019.

Orientador: Ana Maria Frattini Fileti.
Dissertação (mestrado) – Universidade Estadual de Campinas, Faculdade de Engenharia Química.

1. Detectores de vazamento. 2. Gás - Vazamento. 3. Aprendizado de Máquina. I. Fileti, Ana Maria Frattini, 1965-. II. Universidade Estadual de Campinas. Faculdade de Engenharia Química. III. Título.

Informações para Biblioteca Digital

Título em outro idioma: Acoustic method and machine learning algorithms applied to leak detection and localization in low-pressure gas pipelines
Palavras-chave em inglês:
Leak detectors
Gas - Leakage
Machine learning
Área de concentração: Engenharia Química
Titulação: Mestre em Engenharia Química
Banca examinadora:
Ana Maria Frattini Fileti [Orientador]
Roger Josef Zemp
Matheus Souza
Data de defesa: 31-07-2019
Programa de Pós-Graduação: Engenharia Química

Identificação e informações acadêmicas do(a) aluno(a)
- ORCID do autor: https://orcid.org/0000-0001-7230-2364
- Currículo Lattes do autor: http://lattes.cnpq.br/9436934521758192

Folha de Aprovação da Dissertação de Mestrado defendida por Rodolfo Pinheiro da Cruz e aprovada em 31 de julho de 2019 pela banca examinadora constituída pelos doutores:

Profª. Drª. Ana Maria Frattini Fileti (Orientadora) – FEQ/Unicamp

Prof. Dr. Roger Josef Zemp – FEQ/Unicamp

Prof. Dr. Matheus Souza – FEEC/Unicamp

A Ata da Defesa com as respectivas assinaturas dos membros encontra-se no SIGA (Sistema de Fluxo de Dissertação/Tese) e na Secretaria de Pós-Graduação da Faculdade de Engenharia Química – Unicamp.

Ao meu avô Thierry Hugueney (*in memoriam*); aos meus pais Valdir e Deise e a minha namorada Bárbara dedico este trabalho.

Agradecimentos

Agradeço aos meus pais Valdir e Deise pelo apoio e todo esforço por me oferecer uma educação de qualidade.

Aos meus avós Thierry (*in memoriam*) e Carolina pelo incentivo, apoio e por me passarem os valores que me tornaram quem eu sou.

Agradeço a minha namorada Bárbara pela força e estímulo. Obrigado, meu amor, pela paciência.

A minha orientadora Profª. Drª. Ana Maria Frattini Fileti pela confiança e por todo o conhecimento transmitido nesse período.

Ao prof. Dr. Flávio Vasconcelos da Silva por meio do qual tive o primeiro contato com algoritmos de aprendizado de máquina.

Agradeço aos professores da Universidade Federal de Viçosa, Prof. Dr. Andre Gustavo Sato, Prof. Dr. José Vitor Nicácio e Prof. Dr. Erlon Lopez por me estimularem a ingressar na pós-graduação.

Por fim agradeço ao Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq) pelo auxílio financeiro concedido. Agência(s) de Fomento(s) e n°(s) de processo(s): CNPq, 147492/2017-3.

Resumo

Vazamentos de gás colocam em risco vidas humanas e podem provocar sérios desastres ambientais; além disso, são responsáveis por enormes prejuízos econômicos. Em virtude disso, é imprescindível o desenvolvimento de uma técnica de baixo custo que seja capaz de identificar e localizar vazamentos tão logo eles ocorram. O presente trabalho propõe o emprego do método acústico em conjunto com técnicas de aprendizado de máquina para o monitoramento de tubulações de gás de baixas pressões. Busca-se com a técnica proposta uma alternativa para a solução de dois problemas desafiadores para sistemas automáticos de detecção de vazamentos, especialmente quando a tubulação opera sob baixas pressões: a identificação de pequenos vazamentos e a redução de falsos alarmes na presença de distúrbios externos. Essa técnica pressupõe que padrões presentes no comportamento acústico do sistema monitorado podem ser usados para identificar e localizar vazamentos. O aparato experimental empregado consistiu em uma tubulação de cobre (53 m de comprimente e 0,5 polegadas de diâmetro) conectada a um botijão de gás de cozinha alimentado com ar comprimido (100 kPa). Sete orifícios de diferentes diâmetros, variando entre 0,5 mm e 4 mm, foram perfurados ao longo da tubulação, o mais próximo a 0,2 m da entrada da tubulação e o mais afastado a 5,6 m. Cinco microfones, modelo XCM-9767, foram acoplados a estrutura metálica, com distâncias em relação a entrada variando entre 0,1 m e 46,7 m. Quatorze tipos de experimentos foram executados: (1) experimentos sem vazamentos ; (2) sem vazamentos com batidas no botijão;(3) sem vazamentos com batidas na tubulação; (4-10) vazamentos em um único orifício; (11-14) dois vazamentos simultâneos. O comportamento acústico da tubulação nos experimentos foi capturado pelos microfones. O sinal foi convertido do domínio do tempo para o da frequência, ao qual em seguida foram aplicados procedimentos de redução de dimensionalidade. Os dados foram então empregados no treinamento de algoritmos de aprendizado de máquina. Esses modelos foram aplicados tanto para regressão quanto para classificação. Aos algoritmos de classificação coube identificar a qual dos quatorze experimentos uma amostra qualquer apresentada aos modelos pertencia, enquanto os algoritmos de regressão calcularam a posição do orifício a partir do qual o gás estava escapando. Com os dados coletados pelo microfone mais distante dos orifícios atingiu-se uma taxa de detecção dos vazamentos de 99,6%. Tão importante quanto a capacidade de detectar os vazamentos, foi a baixa taxa de falsos alarmes registrada, de apenas 0,3%. Quanto a localização dos vazamentos, o modelo com a melhor

performance atingiu um erro de localização máximo de 4,31%. Os resultados obtidos credenciam a técnica proposta como uma alternativa atrativa para o monitoramento de tubulações de baixa pressão.

Palavras-chave: Detecção de Vazamentos; Localização de Vazamentos; Tubulações de Gás de Baixas Pressões; Algoritmos de Aprendizado de Máquina.

Abstract

Gas leakages are a threat to human life and can cause environmental disasters; moreover, they generate huge economic losses every year. It is indispensable a low-cost technique capable of identifying and locating leaks as soon as they occur. The present work proposes the use of the acoustic method and machine learning algorithms to monitor low-pressure pipelines. It aims to offer an alternative solution for two challenging problems to every automatic leak detection system, mainly when applied to low-pressure pipelines: the detection of small leaks and de reduction of the false alarm rates in the presence of external disturbances. The technique is based on the idea that patterns present in the sound behavior of the system can be used to identify and locate leakages. The experimental apparatus consisted of a copper pipeline (53 m length and 0.5 " diameter) connected to a pressure vessel fed with compressed air (100 kPa). Seven orifices of different diameters, varying from 0.5 mm to 4mm, were drilled along the metallic structure, the closest from the pipeline entrance was 0.2 m, and the furthest was 5.6 m. Five microphones, model XCM -9767, were attached to the pipeline, their distances from the pipeline entrance varying from 0.1 m to 46.7 m. Fourteen types of experiments were conducted: (1) no leakages; (2) no leakages with strikes on the pressure vessel; (3) no leakages with hits on the pipeline; (4-10) one leakage; (11-14) two leakages simultaneously. The microphones captured the acoustic behavior of the system. The signal was converted from the time domain into the frequency domain, to which were then applied two techniques of dimensionality reduction. The data set was fed to some machine learning algorithms. The models were used both for regression and classification. The task of the classification algorithms was to predict to which one of the fourteen experiments a given sample belonged, while the regression ones calculated the position of the orifice from which the gas was leaking. Using the data collected by the microphone that was most distant from the holes, the algorithms were able to identify 99.6% of the leakage. As important as the capacity of detecting the leakages was the low rate of false alarms achieved, only 0.3% of the samples from no leakages experiments were incorrectly classified as leakages. Regarding the localization of the leakages, the model with the best performance achieved a maximum localization error of 4.31%. The results obtained qualify the proposed technique as an appealing alternative to monitor low-pressure pipelines.

Keywords: Lak Detection; Leak localization; Low-Pressure Pipelines, Machine Learning Algorithms.

Lista de Figura

Lista de Tabelas

Abreviaturas

NN - Redes neurais artificiais

FBG - *Fiber Bragg Grating*

A\D - Analógico\Digital

DFT - Transformada discreta de Fourier

FFT - Transformada Rápida de Fourier

J - Unidade Imaginária

KNN – K Vizinhos mais Próximos

RF – Floresta Aleatória

SVM – LINEAR – Máquinas de Vetores Suporte com kernel linear

SVM – RBF – Máquinas de Vetores Suporte com kernel função de base radial

Adaboost – *Adaptative Gradient Boosting*

Xgboost – *Extreme Gradient Boosting*

Sumário

1.1. Introdução

Redes de tubulações são a forma mais eficiente e segura para o deslocamento de fluidos a longas distâncias, apresentando menor taxa de acidentes e danos ambientais que qualquer outra forma de transporte (KENNEDY, 1993). Porém fatores como variações bruscas de pressão, manutenção inadequada, corrosão e defeitos de fabricação podem comprometer esses sistemas e provocar vazamentos, gerando assim prejuízos financeiros em virtude tanto da perda do material transportado quanto da possível interrupção do fornecimento, além da poluição ambiental e riscos à segurança que podem resultar da dispersão do produto no ambiente.

Segundo dados da Petrobras, o número de vazamentos nas tubulações da empresa caiu no período compreendido entre os anos de 2009 e 2013. Mesmo assim, no último ano desse período ainda foram registradas 187 ocorrências. Segundo a companhia esse resultado é superior aos das demais grandes empresas que operam no mercado brasileiro de petróleo e gás (Petrobras, 2014). Jackson et al.(2014) estudaram a ocorrência de vazamentos nas tubulações de gás da cidade americana de Washington. Os autores encontraram um total de 5893 vazamentos; em 12 dos quais a concentração de gás presente na área próxima a tubulação era potencialmente explosiva. Segundo eles a segurança das tubulações americanas aumentou nas décadas recentes, porém incidentes envolvendo vazamentos de gás natural ainda são responsáveis por uma média anual de 17 mortes e prejuízos da ordem de 177 milhões de dólares. Phillips et al.(2013) conduziram trabalho semelhante na cidade de Boston, no qual eles mapearam um total de 3356 vazamentos ao longo de toda a rede de transporte de gás da cidade. Em outro estudo envolvendo gás natural, neste caso focado na análise de ricos em tubulações submarinas, os autores reportaram a ocorrência de 80 incidentes no período de 10 anos entre 1996 e 2006 na região do golfo do México (SLR, 2009). Outro estudo voltado para análise de riscos divulgou a ocorrência de 1326 vazamentos de gás e 108 de petróleo entre os anos de 2001 e 2005 somente na província de Alberta, no Canadá (ALJAROUDI et al., 2015).

Todos esses dados mostram que a ocorrência de vazamentos é um problema comum e recorrente. Em vista disso, torna-se indispensável um método que seja confiável na sua detecção e que também seja capaz de localizá-lo, de modo a possibilitar a rápida correção do problema. Murvay e Silea (2012) descreveram em seu trabalho diversos métodos usados para esse fim, dentre eles o uso de sensores acústicos, balanços de massa

e volume, variações de pressão no escoamento, medições de pressão em pontos da tubulação, métodos baseados em modelagem dinâmica do escoamento etc. Segundo os autores, cada um deles possui vantagens e desvantagens, podendo ser comparados com base em critérios como a sensibilidade para detectar pequenos vazamentos, o custo, capacidade de estimar a localização, a necessidade de manutenção e taxa de falsos alarmes.

O sistema estudado neste trabalho consistiu em uma tubulação de gás operando sob baixa pressão (100 kPa). Dentre as tubulações que operam sob baixas pressões estão os sistemas de distribuição direta de gás aos consumidores tanto em residências quanto em pontos comerciais. Esses sistemas são cômodos e têm diversas vantagens em relação ao tradicional de compra e armazenamento de botijões de gás, entre elas o fornecimento ininterrupto, a eliminação dos riscos relacionados à estocagem, a liberação de espaço físico, a possibilidade do gás ser usado tanto no preparo de alimentos quanto no aquecimento de água para o banho e o modelo de pagamento no qual só é paga a quantidade utilizada no intervalo de tempo considerado, de forma semelhante à conta de energia elétrica (ROSA, 2016). Porém o sistema traz riscos, uma vez que o gás natural é inflamável e vazamentos podem resultar em explosões. Exige-se assim uma total confiabilidade operacional das tubulações.

Para o monitoramento do sistema foram empregados microfones. Essa é uma técnica intrusiva, que apesar dos inconvenientes, não apresenta diversas restrições enfrentadas pelas técnicas não intrusivas, como a elevada atenuação do sinal que pode ser provocada pelo ambiente que envolve a tubulação (tipo de solo em que a tubulação esteja enterrada por exemplo), pelo tipo de material e diâmetro da tubulação (o sinal tende a se dissipar rapidamente na medida em que o diâmetro da tubulação aumenta e em tubulações de plástico), e também em tubulações que operam sob baixas pressões (KHALIFA et al., 2010). Outra vantagem no uso dos microfones é o comportamento do sinal capturado quando se inicia um vazamento, ele apresenta um pico e atinge um novo estado estacionário diferente do inicial. Esse fato reduz a taxa de falsos alarmes, uma vez que facilita a distinção entre o sinal gerado por um vazamento e o causado por uma perturbação externa qualquer.

O sinal acústico gerado na operação da tubulação em diversas condições foi captado pelos microfones e usado no treinamento de algoritmos de aprendizado de máquina: Esses algoritmos são capazes de identificar automaticamente padrões e aprender a partir dos dados que lhe são fornecidos. Esses padrões presentes no sinal

acústico permitiram detectar e localizar os vazamentos provocados nos experimentos. O sucesso alcançado nas duas tarefas qualifica a técnica proposta no presente trabalho como uma alternativa atrativa para o monitoramento de tubulações de baixa pressão.

1.2. Hipóteses

1. O sinal acústico provocado pelo surgimento de vazamentos em uma tubulação pode ser captado por microfones e usado como instrumento para a detecção e localização dos vazamentos;
2. Algoritmos de aprendizado de máquina são uma técnica eficiente para identificar os padrões presentes nos sinais captados pelos microfones;
3. A redução da dimensionalidade dos dados facilitará o treinamento dos algoritmos e trará melhoras no desempenho;
4. Os padrões presentes nos sinais captados permitirão detectar e localizar vazamentos.

1.3. Objetivos

Dada a importância da confiabilidade operacional em redes de tubulações, busca-se nesse trabalho estudar a viabilidade e a eficiência de um sistema de detecção e localização de vazamentos baseado no uso de microfones e algoritmos de aprendizado de máquina.

1.3.1. Objetivos Específicos

- Obtenção dos dados experimentais para verificação de possíveis assinaturas típicas das situações com e sem a ocorrência de vazamentos;
- Utilização de técnicas de processamento de sinais tais como transformada rápida de Fourier, Análise de Componentes Principais e algoritmos de aprendizado de máquina;
- Encontrar os algoritmos com as melhores performances nas tarefas de identificar e localizar os vazamentos;
- Avaliar as predições dos modelos quando lhes são apresentados dados completamente diferentes dos fornecidos durante o treinamento.

2. Revisão Bibliográfica

Neste capítulo serão descritos trabalhos recentes desenvolvidos com a finalidade de identificar e localizar vazamentos. Essas técnicas incluem balanço de massa e volume, método transitório de pressão, modelagem em tempo real, métodos estatísticos, análise pontual de pressão, métodos ópticos, ultrassônicos e acústicos.

2.1. Balanço de Massa/Volume

Técnicas de detecção de vazamentos que empregam balanços de massa e de volume baseiam-se no princípio de conservação de massa. Os balanços podem ser calculados com o uso de medições de vazão, temperatura e pressão do escoamento. Vazamentos são detectados quando ocorre uma discrepância superior a determinado limite estipulado entre as medidas de volume na entrada e na saída de determinado segmento de uma tubulação. Esse limite é estabelecido, ou seja, o volume na entrada não é exatamente igual ao da saída em virtude de alterações que podem ser provocadas por mudanças de temperatura e pressão (ADEC., 1999).

Segundo Murvay e Silea (2012) as vantagens desse conjunto de técnicas incluem seu baixo custo e a facilidade tanto operacional quanto de instalação dos equipamentos. Além disso, com esses métodos é possível a detecção de pequenos vazamentos, embora para isso seja necessário longo intervalo de tempo, podendo chegar a 60 minutos na detecção de vazamentos da ordem de 1% do material transportado (DOORHY, 2011). Porém, entre suas limitações estão à impossibilidade de se localizar o vazamento e a não aplicabilidade durante períodos de operação transiente (MURVAY; SILEA, 2012).

2.2. Método Transitório de Pressão

Essa técnica baseia-se na análise de variações de pressão no interior das tubulações. Um vazamento provoca uma repentina queda de pressão no local onde ocorreu a ruptura, a partir da qual é gerada uma onda que se propaga pelo fluido a velocidade do som em ambas as direções. Transdutores de pressão instalados ao longo da tubulação captam essas variações e através da análise e processamento do sinal captado busca-se identificar se essa queda de pressão foi provocada ou não por um vazamento. Além disso, a diferença no intervalo de tempo para que as ondas atinjam os sensores posicionados em lados opostos ao vazamento permite identificar a sua localização (MURVAY; SILEA, 2012).

Diferentes tipos de sensores e métodos de tratamento de dados podem ser aplicados nessa técnica. Hou et al. (2013) empregaram em seu trabalho sensores de fibra óptica (*Fiber Bragg Grating* (FBG)) acoplados à parede externa da tubulação para captar as variações de pressão. Os sensores FBG consistem em uma fibra óptica com filtros em seu interior. Parte da luz no interior das fibras é refletida por esses filtros. O comprimento das ondas refletidas varia com a temperatura e a deformação da fibra. Dessa forma o sensor é calibrado para indicar a temperatura ou a pressão a que ele está submetido de acordo com o comprimento de onda que é refletido. O princípio de funcionamento é mostrado na Figura 1, enquanto na Figura 2 estão alguns modelos que empregam a tecnologia disponíveis comercialmente. Segundo os autores esse tipo de sensor tem como vantagem a alta sensibilidade, confiabilidade no longo prazo, imunidade a interferências eletromagnéticas, além de serem pequenos, leves e de fácil instalação. Para aumentar a precisão na determinação da localização do vazamento os autores consideraram nos cálculos a variação da velocidade tanto das ondas de pressão quanto do escoamento do fluido. Já para o tratamento dos dados coletados eles aplicaram a transformada Wavelet para localizar os pontos de queda de pressão correspondentes aos vazamentos, obtendo por fim um erro absoluto máximo de 0,38 m, o que demonstrou à boa acurácia do método proposto.

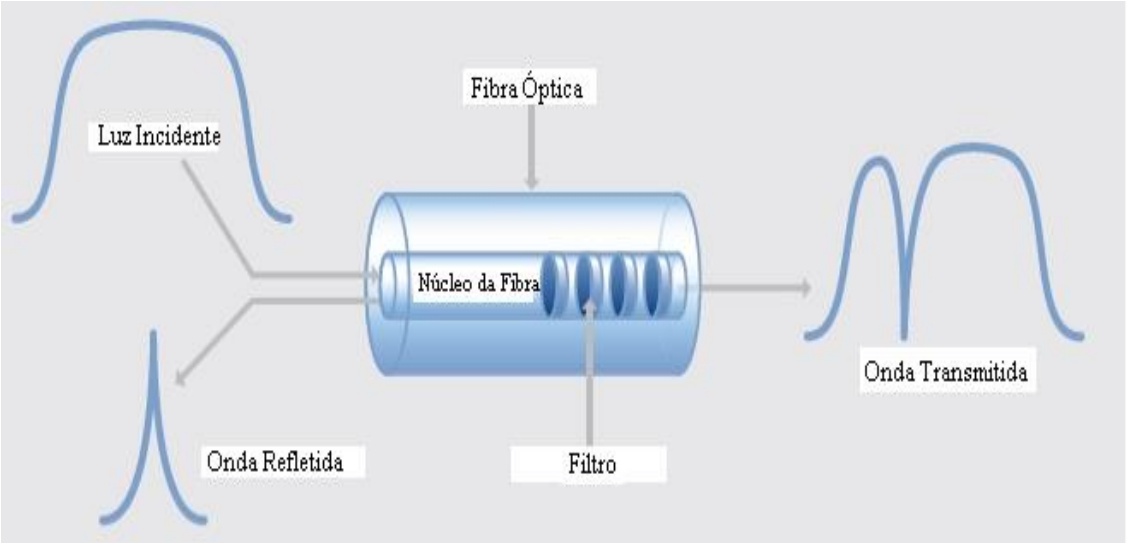

Figura 1 - Princípio de funcionamento de um sensor FBG. Fonte: Adaptado de National Instruments (2017).

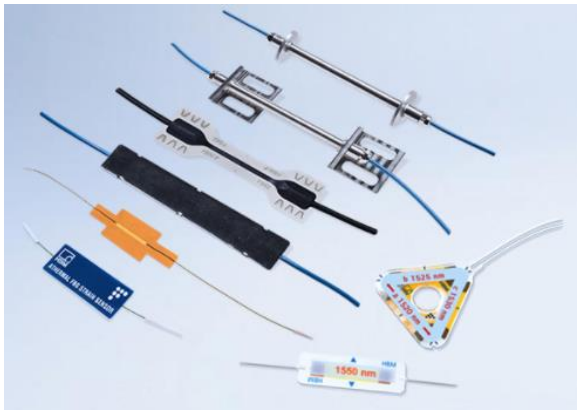

Figura 2 - Diversos modelos de sensores que empregam a tecnologia FBG. Fonte: HBM Fiber Sensing.

Essa técnica é especialmente adequada para a rápida detecção de grandes vazamentos, em virtude da alta velocidade de propagação das ondas de pressão nos fluidos transportados, porém ela pode falhar na detecção de pequenos vazamentos (ADEC., 1999). Segundo Abdulshaheed; Mustapha e Ghavamian (2017) ela é mais indicada para tubulações transportando líquidos, em virtude da maior perda de energia das ondas de pressão nos fluidos gasosos. Outra desvantagem desse método é a sua sensibilidade a mudanças operacionais, como por exemplo, a abertura de uma válvula, e também a presença de ruídos, uma vez que mesmo pequenas variações de pressão podem gerar falsos alarmes (ABDULSHAHEED; MUSTAPHA; GHAVAMIAN, 2017). Buscando superar essa limitação Souza e Hoffman (2011) desenvolveram um sistema baseado na análise de mapas de pressão, no qual diversas características da onda de pressão são analisadas durante a sua propagação pela tubulação. Três algoritmos são responsáveis por filtrar os ruídos e montar o mapa tridimensional, permitindo assim a identificação dos eventos provocados por vazamentos.

2.3. Modelagem em Tempo Real (RTTM)

Este é entre todos os métodos o que apresenta a maior sensibilidade, porém é caro e complexo, uma vez que requer grande quantidade de instrumentos para monitoramento em tempo real das condições do escoamento na tubulação (MURVAY; SILEA, 2012). Os modelos são criados com base nos princípios de conservação de massa, momento e energia, e descrevem o comportamento do escoamento na tubulação tanto em regime transiente quanto no estado estacionário. São necessárias medidas de pressão, temperatura e vazão. Os vazamentos são identificados quando ocorre uma diferença entre os valores medidos e os calculados pelos modelos (MURVAY; SILEA, 2012).

O sucesso desse método depende da qualidade e da quantidade dos sensores empregados. Erros de calibração podem resultar em falsos alarmes ou na não detecção de um vazamento, além da perda de determinado sensor em certas situações exigir o desligamento de todo o sistema (ADEC., 1999). Buscando resolver esses problemas Lu; She e Loewen (2017) estudaram o efeito das incertezas presentes nas medidas dos sensores, nos sistemas de aquisição de dados e nas informações relativas as propriedades dos fluidos na capacidade de um sistema RTTM em localizar vazamentos, fornecendo medidas quantitativas da influência de cada uma delas no desempenho do sistema. Esse tipo de medida possibilita a sintonização do modelo para a distinção entre erros instrumentais e outros distúrbios da ocorrência de vazamentos.

Para reduzir o custo computacional desse método Ruiz; Mujica e Mujía (2015) desenvolveram um método que combinou RTTM com análise estatística. Os autores construíram perfis de pressão ao longo de uma tubulação em diversas simulações e utilizaram a ferramenta estatística análise de componentes principais (PCA) para determinar as informações mais relevantes. Os resultados encontrados mostraram a possibilidade de se detectar e até mesmo localizar o vazamento comparando-se os dados de pressão coletados durante a operação com os mapas construídos.

A grande vantagem da modelagem em tempo real é a sua capacidade de englobar nos modelos todas as características do escoamento do fluido (temperatura, vazão e pressão), além de levar em conta também a configuração física da tubulação (comprimento, material, diâmetro, etc.) e as propriedades do fluido transportado (ADEC., 1999). Porém não se deve perder de vista que o aumento da complexidade do modelo aumenta os custos computacionais envolvidos.

2.4. Métodos Estatísticos

Nesse método análises estatísticas são realizadas em alguma variável do escoamento. Para isso é necessária uma prévia sintonização para se determinar o comportamento da variável sob diferentes condições operacionais na ausência de vazamentos. Durante o monitoramento, medidas da variável são tomadas em diversos pontos da tubulação. Os dados obtidos são agrupados com o uso de técnicas de clusterização, que são formas não supervisionadas de classificação de dados, nas quais eles são reunidos em *clusters* de acordo com alguma medida de similaridade (JAIN; MURTY; FLYNN, 1999). O vazamento é detectado comparando-se os dados coletados continuamente com os padrões obtidos pela clusterização (MURVAY; SILEA, 2012).

As variáveis empregadas costumam ser a vazão e a pressão. Segundo Zhang et al. (2014) o balanço estatístico de volume vem sendo empregado com sucesso em mais de 600 tubulações ao redor do mundo, das mais variadas dimensões e transportando diversos tipos de fluidos. Pesquisando essa técnica, Di Blasi e Muravchik (2009) ampliaram o uso da estatística, empregando suas ferramentas também na determinação da quantidade de matéria acumulada no interior da tubulação, termo que entra nos balanços de volume, simplificando sobremaneira os cálculos necessários.

Ferramentas estatísticas também vêm sendo aplicadas em conjunto com redes neurais artificiais. As primeiras são usadas para o tratamento dos dados experimentais a fim de reduzir o número de informações usadas como entrada na etapa de treinamento das redes. Essa prática elimina os dados que não trazem informações relevantes e diminui a complexidade do treinamento. Santos (2015) e Fernandes (2017) usaram a ferramenta Análise de Componentes Principais (PCA) para tratar dados coletados por sensores acústicos. Zadkarami; Shahbazian e Salahshoor (2016) também trabalharam com ferramentas estatísticas e redes neurais. Os autores não realizaram experimentos, os dados de pressão e vazão da tubulação foram obtidos por meio de simulações. A combinação dessas duas técnicas permite não somente a detecção dos vazamentos, mas também estimar a sua localização e magnitude.

De modo geral as grandes vantagens das técnicas estatísticas residem na sua capacidade de operar sob regimes transientes, a baixa taxa de falsos alarmes e a elevada sensibilidade (ZHANG et al., 2014).

2.5. Análise Pontual de Pressão

Essa técnica baseia-se na queda de pressão provocada no interior da tubulação quando ocorre um vazamento. Medidas são tomadas continuamente em diversos pontos e, com o uso de ferramentas estatísticas, identifica-se a ocorrência do vazamento com base na diferença dos padrões das medidas na presença e na ausência de vazamentos (MURVAY; SILEA, 2012).

A grande desvantagem desse método é a sua elevada sensibilidade à presença de ruídos. Alterações operacionais que provocam mudanças de pressão podem gerar alarmes falsos. O baixo custo e a sua simplicidade o tornam vantajoso para tubulações pequenas que operam em regime permanente (SCOTT, S; BARRUFET, 2003).

2.6. Métodos Ópticos

Os métodos ópticos podem ser divididos em duas categorias, os ativos e os passivos. Os primeiros empregam fontes emissoras de radiação, os vazamentos são detectados quando ocorre espalhamento ou absorção anormal da radiação emitida. Já os segundos não utilizam as fontes emissoras, eles captam a radiação emitida pelos fluidos (MURVAY; SILEA, 2012).

As técnicas básicas para monitoramento ativo incluem diodos laser ajustáveis de absorção espectroscópica (*Tunable Diode Laser Absorption Spectroscopy* TDLAS), Laser induzindo Fluorescência (LIF), Espectroscopia anti-raman coerente (CARS), Transformada de Fourier de espectroscopia infravermelha (FTIR) e Sensoriamento evanescente (MURVAY; SILEA, 2012) .Em relação a elas, o grande desafio tecnológico é encontrar uma fonte emissora de baixo custo e que seja capaz de detectar os vazamentos a distâncias longas. Suas vantagens são a portabilidade e elevada sensibilidade, qualidades também presentes nos métodos passivos. Nesse sentido Jobs et al. (2016) desenvolveram um sistema capaz de detectar concentrações de N_2O e CH_4 da ordem de 2,6 e 0,4 ppm respectivamente, porém o limite máximo de detecção alcançado foi de 40 m.

Os cabos de fibra óptica estão entre os métodos ativos empregados. Eles devem ser instalados ao longo de toda a tubulação, conforme pode ser visto na Figura 3. Os vazamentos são detectados pelas mudanças de temperatura provocadas pela presença do fluido que vazou. Essa variação altera os índices de refração dos cabos. Emitindo-se pulsos de laser e analisando-se o seu comportamento é então possível detectar tanto o vazamento quanto a sua localização (GEIGER, 2006). Essa técnica tem a vantagem de ser rápida na detecção e tem também alta sensibilidade. Inaudi e Bont (2013) conseguiram detectar a ocorrência dos vazamentos em cerca de 20 s a 30 s após o início dos mesmos. Segundo os autores essa é uma técnica especialmente adequada para tubulações transportando amônia, já que é uma substância que provoca rápidas e grandes variações de temperatura ao se evaporar. Essa rapidez de detecção também é demonstrada no trabalho de Wang et al. (2017). A nova tecnologia empregada pelos autores foi capaz de captar variações de temperatura da ordem de 0,0005 °C, além disso, ela é capaz também de captar sinais acústicos, aumentando assim a confiabilidade do sistema. A aplicação de fibras ópticas para a captação de sinais acústicos é mostrada esquematicamente na Figura 4. Esse esquema mostra o pulso luminoso emitido, a presença do sinal acústico provoca

uma deformação microscópica na fibra, que resulta na mudança da relação de fase entre a luz emitida e a refletida. Essa mudança é captada pelo detector e permite identificar a presença do sinal acústico. Apesar das vantagens, o uso das fibras ópticas esbarra em seu elevado custo e também na dificuldade de instalação em tubulações enterradas já existentes, uma vez que nesse caso é necessário perfurar o solo para posicionar os cabos próximos à tubulação (MURVAY; SILEA, 2012).

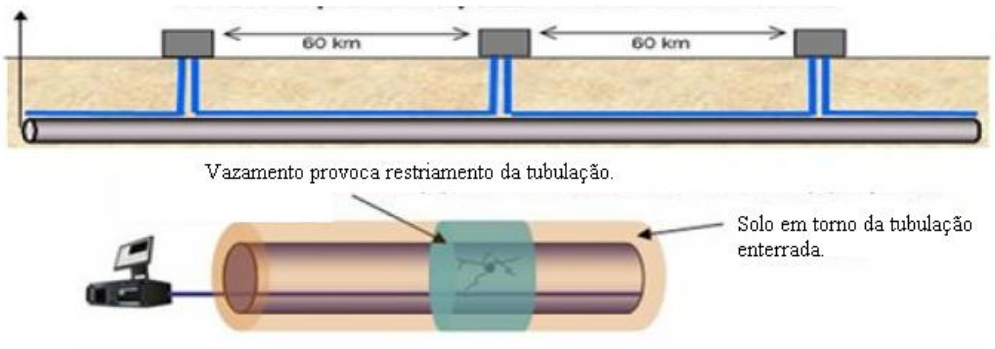

Figura 3 - Instalação de cabo de fibra óptica para a detecção de vazamentos em uma tubulação enterrada. Fonte: Adaptado de Geiger (2006)

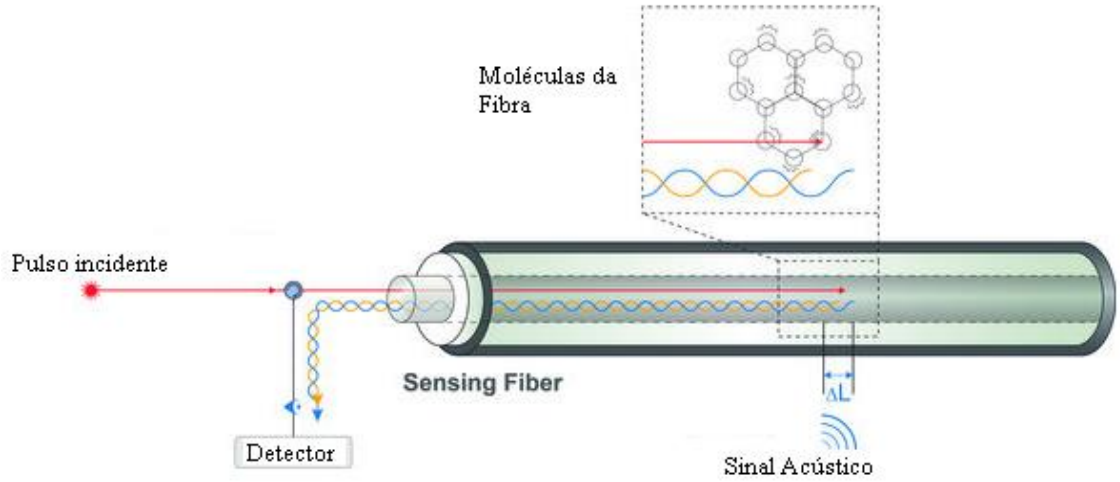

Figura 4 - Tecnologia para a detecção de sinais acústicos através de fibras ópticas. Fonte: Adaptado de AP Sensing.

Outros tipos de cabos podem ser empregados, mas o princípio de funcionamento não se baseia mais nas mudanças de propriedade ópticas. Cabos elétricos podem ser construídos com materiais que tenham alguma de suas propriedades, como por exemplo a impedância e a resistência, alteradas ao entrarem em contato com o fluido transportado. Essa alteração é usada para a detecção dos vazamentos (ADEC., 1999).

Os custos dos sistemas passivos são menores em relação aos ativos devido à ausência da fonte emissora de radiação, mas ainda assim continuam altos, em virtude da necessidade de gravadores e receptores de imagens. Dessa técnica os sistemas mais

empregados, especialmente na localização de vazamentos em dutos de gás, são o imageamento térmico e o imageamento multionda. Os sensores costumam ser acoplados a veículos em movimento. Os helicópteros são os mais empregados para esse fim pela capacidade de voar próximo as tubulações e a pequenas altitudes. Ershov; Klimov e Vavilov (2007) desenvolveram um sistema para o monitoramento de gasodutos no qual o equipamento, carregado por um helicóptero, combinava três métodos detecção, um método óptico ativo com fonte emissora de lasers diodos e duas passivas. O sistema desenvolvido tem elevada sensibilidade, com capacidade de detecção de gases na concentração de 150 ppm para um alcance máximo de 150 m. Porém os custos são obviamente elevados.

2.7. Sensores Ultrassônicos

Sensores ultrassônicos também podem ser empregados na detecção de vazamentos. Com eles é possível determinar a velocidade do escoamento no interior da tubulação e a partir dela a vazão volumétrica. Com esses dados calcula-se o balanço de volume em segmentos da tubulação para a identificação dos possíveis vazamentos. Segundo Santos e Younis (2011) duas técnicas ultrassônicas são adequadas para esse fim, uma baseada no efeito Doppler e outra no tempo de trânsito da onda ultrassônica emitida entre dois transdutores. Os autores desenvolveram um sistema que combinou as duas, tornando-a aplicável a uma ampla gama de diferentes fluidos. Os transdutores acoplados a tubulação enviavam dados a um computador por meio de rede wireless, permitindo assim o acompanhamento em tempo real. Os autores mostraram a capacidade do sistema para o cálculo das vazões, mas não implementaram os algoritmos para detecção dos vazamentos. Já Dudić et al. (2012) compararam o desempenho de um sistema ultrassônico e de outro óptico baseado em imageamento térmico na detecção e na quantificação de vazamentos em um sistema experimental transportando ar comprimido. Nos experimentos foram provocados vazamentos de diferentes magnitudes, e em todos eles o sistema ultrassônico foi capaz de detectá-los, porém em relação a quantificação ele funcionou somente para os vazamentos de magnitudes menores. O problema desse método foi a queda de seu desempenho com a presença de ruídos externos. Já com o sistema óptico não foi possível detectar os vazamentos menores, uma vez que a diferença de temperatura provocada pelo do ar liberado foi pequena. Guenther e Kroll (2016) pesquisaram um sistema baseado na análise dos sinais ultrassônicos, no qual a ocorrência do vazamento é detectada com base no comportamento do sinal captado. Esse sistema

dispensa os cálculos de balanço de volume, porém segundo os autores, ele muito sensível a ruídos externos e ainda deve ser testado em um ambiente industrial.

2.8. Sensores Acústicos

As ondas sonoras geradas por um vazamento podem ser usadas para detectá-lo, determinar sua localização e magnitude. O princípio de dessa técnica é a comparação do comportamento das ondas sonoras em situações operacionais normais e com o aparecimento de vazamentos (ADNAN et al., 2015).

Três tipos de sensores são comumente aplicados para a captura do sinal acústico: microfones, acelerômetros e sensores de pressão dinâmica. O sinal captado pelos dois primeiros se dissipa mais rapidamente, esse problema é minimizado no caso dos microfones pelo baixo custo dos sensores. Já no caso dos sensores de pressão dinâmica o sinal pode se propagar por maiores distâncias, razão pela qual a maioria das pesquisas recentes se concentram na aplicação desse tipo de sensor (LIU et al., 2017a).

Inúmeras alternativas já foram propostas para o tratamento do sinal capturado pelos sensores. MENG et al. (2012) empregaram sensores de pressão dinâmica para o monitoramento de uma tubulação transportando gás. Para a identificação dos vazamentos os autores calcularam três características do sinal captado na tubulação em experimentos com vazamentos e em situações operacionais normais. As três características calculadas foram o somatório acumulado das diferenças, o valor médio e o valor máximo, todas elas computadas considerando-se um mesmo intervalo de tempo. Os vazamentos foram identificados através da comparação dos valores dessas três características nas duas situações (vazamentos e sem vazamentos). Segundo os autores a comparação simultânea das três características contribuiu para reduzir a taxa de falsos alarmes.

O método acústico mais empregado para a localização dos vazamentos consiste em se posicionar dois sensores nos extremos opostos da seção da tubulação monitorada. Mede-se os intervalos de tempo necessários para o sinal provocado por um vazamento atingir os dois sensores (Figura 5). O valor da diferença entre esses dois intervalos é então aplicado na Equação (1) para o cálculo da posição, na qual X é a posição do vazamento, L é a distância entre os dois sensores, α é a velocidade de propagação do sinal e Δt é a diferença entre os dois intervalos (conforme esquema da Figura 5) (MENG et al., 2012). Os resultados obtidos através dessa técnica podem ser aprimorados com a instalação de dois sensores de um mesmo lado da tubulação separados por uma pequena distância. A medição da diferença de tempo pra atingir esses dois sensores é também empregada para

cálculo da localização do vazamento (LIU et al., 2019b). A principal dificuldade no uso das técnicas que empregam a diferença nos intervalos de tempo para o sinal atingir diferentes sensores está na determinação da velocidade de propagação do sinal, uma vez que ela depende das condições operacionais a que a tubulação está submetida. Novos métodos para o cálculo dessa velocidade foram desenvolvidos de forma a incluir nos cálculos a influência da temperatura, pressão, densidade (MENG et al., 2012; JIN et al., 2014)

$$X=\frac{L-\alpha\Delta t}{2}, \tag{1}$$

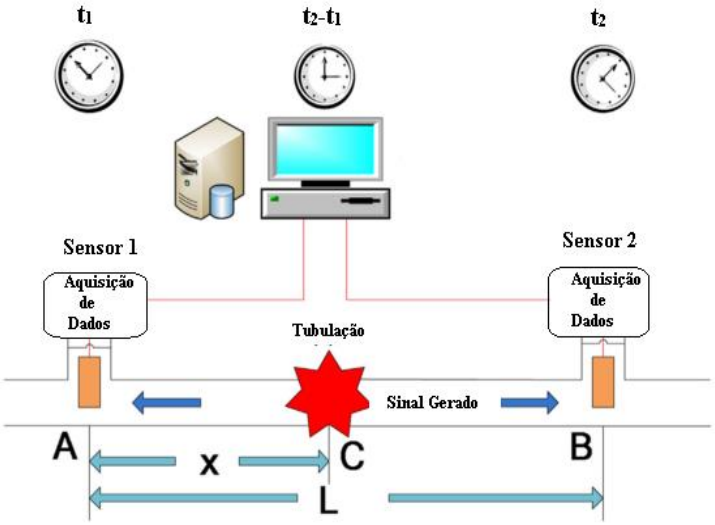

Figura 5 - Método de detecção de vazamentos baseado no intervalo de tempo que o sinal acústico gerado leva para atingir diferentes sensores .Fonte: MENG et al. (2012).

Outra técnica acústica para a localização de vazamentos emprega modelos de propagação das ondas acústicas. O método requer a medida da amplitude do sinal acústico, fugindo-se assim das incertezas presentes no cálculo da velocidade de propagação do sinal e da determinação das diferenças nos intervalos de tempo para se atingir dois sensores. A amplitude medida é empregada em equações que descrevem a propagação do sinal acústico, assim a magnitude da atenuação do sinal indicará a localização do vazamento. O uso dessa técnica levou a menores erros no comparativo com os métodos descritos anteriormente que empregam os intervalos de tempo (CUI-WEI et al., 2015; LIU et al., 2019a). Porém é necessária a obtenção de um modelo que descreva com exatidão a propagação das ondas, o que é difícil dado o complexo comportamento do escoamento dos gases na tubulação.

Uma desvantagem do método acústico é a elevada presença de ruídos no sinal capturado pelos sensores, o que dificulta a obtenção dos parâmetros necessários para a

identificação e localização dos vazamentos. Em vista disso grandes esforços vem sendo direcionados na busca de métodos de permitam separar com eficiência o sinal acústico dos ruídos (MOSTAFAPOUR; DAVOODI; GHAREAGHAJI, 2014; LIU et al., 2019b)

Uma alternativa aos métodos já descritos emprega algoritmos de aprendizado de máquina. Esses algoritmos identificam automaticamente padrões pressentes no sinal acústico. As diferenças nesses padrões são usadas tanto para identificar quanto para localizar vazamentos (SANTOS et al., 2014; EL-ZAHAB et al., 2018). Com o uso desses algoritmos são necessárias somente medidas da amplitude do sinal. Além disso, evita-se a árdua tarefa de desenvolvimento de modelos para o cálculo da velocidade do sinal ou que descrevam a propagação do sinal acústico.

2.9. Comparativo

Os métodos de detecção apresentados variam em suas propriedades físicas e operacionais. Nenhum deles é universalmente aplicável e apresenta todas as qualidades desejáveis. Assim o uso conjunto de diferentes técnicas supera limitações de um sistema operando individualmente, evitando assim falsos alarmes e que vazamentos passem despercebidos.

A Tabela 1 apresenta um comparativo dos custos, velocidade de detecção, facilidade de uso, capacidade de localização, determinação da magnitude dos vazamentos e sensibilidade dos métodos descritos. As informações quanto ao custo devem ser vistas com cautela, uma vez que os autores não especificaram quais são limites usados para a classificação em cada uma das categorias. Por exemplo, eles classificaram o método de modelagem (RTTM) e os acústicos como de custo elevado. O primeiro exige instrumentação para medidas de vazão, pressão e temperatura, enquanto o segundo somente microfones (outros sensores acústicos podem ser usados), que são sensores de baixo custo. Mesmo a exigência de diversos microfones ao longo de toda a extensão da tubulação não torna próximos os custos desses dois métodos. Problema semelhante ocorre na comparação entre os métodos ópticos e os acústicos. Outro problema está no comparativo entre a sensibilidade das técnicas, uma vez que autores diferentes tendem a usar métricas diferentes para quantificar essa variável.

As informações contidas na Tabela 1 confirmam a afirmação de que nenhum método possui todas as qualidades desejáveis. Os métodos com as maiores sensibilidades (capacidade de detectar pequenos vazamentos) são os mais caros e difíceis de serem

implementados, enquanto o contrário vale para os métodos de menor sensibilidade. Assim a pergunta quanto ao melhor método só pode ser respondida tendo em vista a aplicação a que ele se destina.

Tabela 1 - Comparativo entre métodos de detecção de vazamentos descritos. Fonte: Adaptado de Murvay e Silea (2012).

	Característica					
Método	Custo	Velocidade Detecção	Facilidade Operacional	Estimativa Localização	Estimativa Magnitude	Sensibilidade
Balanços de Massa/ Volume	Baixo	Baixa	Sim	Não	Sim	Baixa
Transitório Pressão	Baixo	Rápida	Sim	Sim	Sim	Baixa
Modelagem	Alto	Rápida	Não	Sim	Sim	Alta
Estatísticos	Alto	Rápida	Sim	Sim	Sim	Variável de acordo com o sensor
Análise Pontual de Pressão	Baixo	Rápida	Sim	Não	Não	Baixa
Ópticos	Alto	Média	Sim	Sim	Sim	Alta
Fibra Óptica	Alto	Rápida	Sim	Sim	Sim	Alta
Ultrassônicos	Alto	Rápida	Sim	Sim	Não	Alta
Acústicos	Alto	Rápida	Sim	Sim	Sim	Variável de acordo com o sensor

2.10.Conclusão

Conforme exposto existem diversas técnicas disponíveis para o monitoramento de tubulações de gás. Cada uma delas tem suas próprias vantagens e desvantagens. De modo geral os métodos com maior sensibilidade e confiabilidade tem custos proibitivos. Pesquisas vem sendo conduzidas no sentido de aprimorar as técnicas já existentes, ao mesmo tempo em que são propostas novas metodologias.

Dentre todos os métodos o acústico se mostra o mais promissor (LIU et al., 2017a). Essa técnica tem diversas características positivas. Além disso, constantes progressos vêm sendo obtidos, existindo ainda muitas possibilidades de aprimoramento. Ela pode ser aplicada nas mais diversas condições operacionais, a depender do tipo de sensor empregado.

Neste trabalho propõe-se uma metodologia para monitorar tubulações de gás que operam sob baixas pressões, situação especialmente desafiadora e que requer sensores com elevada sensibilidade, o que motivou a escolhas dos microfones. A grande maioria dos trabalhos disponíveis na literatura tratam do monitoramento de tubulações que operam sob altas pressões. Um exemplo desse tipo são as tubulações que transportam gás natural por longas distância, das regiões produtoras até as consumidoras. Entretanto, as tubulações de gás de baixas pressões estão presentes em diversas aplicações. As redes de distribuição de gás aos consumidores nas cidades operam sob baixas pressões. Além disso, elas também estão presentes em todo tipo de ambiente industrial.

Para o tratamento do sinal captado pelos microfones propõe-se o uso de técnicas estatísticas como análise de componentes principais e algoritmos de aprendizado de máquina.Com elas busca-se encontrar padrões permanentes presentes no sinal gerado por vazamentos que permitam localizá-los e identificá-los. O método mais comumente empregado consiste em calcular algum tipo de característica do sinal acústico, porém as variações nessas características geradas por vazamentos também podem ser provocadas por outros tipos de perturbações na tubulação, o que acaba por gerar falsos alarmes (sistema de detecção acusa ocorrência de vazamento durante operação normal).

Sendo assim, busca-se com a metodologia proposta neste trabalho contribuir no aprimoramento do método acústico, buscando uma técnica eficaz e confiável no monitoramento de tubulações que operam sob baixas- pressões.

3.Fundamentação Teórica

Nesta seção serão tratados conceitos a respeito do processamento digital de sinais, da transformada de Fourier e dos algoritmos de aprendizado de máquina empregados.

3.1. Processamento de Sinais

Sensores e transdutores são instrumentos responsáveis por transformar as grandezas físicas do mundo real em sinais elétricos. Esse sinal elétrico tem natureza analógica, contínua no tempo. Para trazer a grandeza física para dentro do computador ou qualquer outro dispositivo digital é necessário convertê-la para linguagem de máquina, de forma que possa ser tratado e processado digitalmente. O equipamento responsável por essa transformação é conhecido como conversor A/D (Analógico/Digital).

Para a digitalização de um sinal analógico são necessárias três etapas: amostragem, quantização e codificação. A primeira é a medição da amplitude do sinal em instantes discretos, ou seja, é a obtenção de amostras do sinal a cada intervalo de tempo constante definido. Esse intervalo de tempo é a periodicidade de amostragem e seu inverso é a frequência de amostragem. De acordo com o teorema da amostragem de Nyquist (SMITH, 1997), a frequência de amostragem deve ser maior ou igual a duas vezes a maior frequência do sinal amostrado, caso contrário ocorrerá um fenômeno conhecido como efeito aliasing, que gera distorção do sinal devido a uma taxa insuficiente de amostragem. Essa taxa insuficiente acaba por tornar indistinguíveis diferentes sinais amostrados. A Figura 6 mostra um exemplo do efeito aliasing, a linha contínua é o sinal analógico e os pontos o sinal digitalizado, na Figura 6 (a) a frequência da amostragem escolhida obedeceu ao critério de Nyquist, o que não ocorreu na Figura 6 (b), onde pode-se ver claramente a diferença entre o sinal analógico e o digitalizado, resultando assim na perda da informação contida no sinal original.

A etapa de quantização consiste em atribuir valores, pertencentes a um conjunto finito de valores possíveis chamado níveis de quantização, a cada um dos valores amostrados. Cada amplitude é colocada no nível de quantização mais próximo, de forma a minimizar o erro absoluto da conversão. O tamanho desse conjunto de valores possíveis depende do número de bits do conversor (A/D) empregado. Para n bits, serão no total 2^n níveis de quantização. Assim quanto maior o número de bits mais exata é a conversão

A última etapa do processo é a codificação, na qual são atribuídos números binários a cada um dos níveis de quantização. O número binário obtido ao final é o sinal digitalizado.

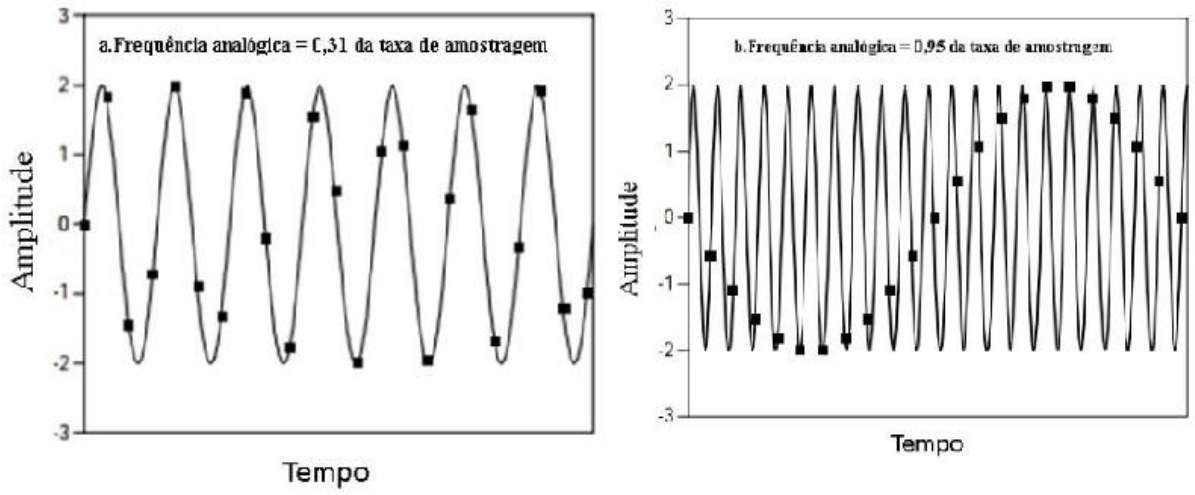

Figura 6 - Sinais analógicos digitalizados com diferentes frequências de amostragem. Em (a) a frequência usada obedeceu ao critério de Nyquist, o que não ocorreu em (b).Fonte: Adaptado de Smith (1997)

3.2. Transformada de Fourier

A transformada de Fourier é aplicada em diversos campos da engenharia. Ela é um dos procedimentos mais comuns e importantes na área do processamento digital de sinais. Através dela um sinal no domínio do tempo é convertido para o domínio da frequência, o que torna mais fácil e esclarecedora a sua análise e manipulação (LYONS, 2011).

Essa técnica matemática decompõe um sinal em suas componentes elementares de senos e cossenos. Qualquer outra função poderia ser usada na decomposição, os senos e cossenos são empregados devido a uma propriedade, não compartilhada pelo sinal original, que garante que um sinal senoidal na entrada de um sistema linear e invariante no tempo mantem essa característica na saída, somente sua fase e amplitude podem variar (SMITH, 1997).

A transformada discreta de Fourier (DFT) é a ferramenta usada para lidar com sinais discretos. Neste trabalho o sinal discreto é um conjunto de valores obtido da amostragem periódica de um sinal no domínio do tempo, assim, usando-se esse sinal como entrada, a DFT dará como resposta o mesmo sinal decomposto no domínio da frequência.

A DFT surgiu da transformada contínua de Fourier, definida como:

$$X(f)= \int_{-\infty}^{+\infty} x(t)e^{-j2\pi ft}dt, \quad (2)$$

na qual X(f) é o sinal no domínio da frequência e x(t) é o sinal contínuo no domínio do tempo. Com o desenvolvimento dos computadores e a necessidade de se lidar com os sinais digitais criou-se a DFT, expressa por

$$X(m)= \sum_{n=0}^{N-1} x(n)e^{-j2\pi nm/N}, \quad (3)$$

que com o uso da relação de Euler $e^{-j\alpha}= \cos\alpha - j\sin\alpha$ pode ser escrita de forma equivalente como

$$X(m) = \sum_{n=0}^{N-1} x(n)[cos(2\pi nm/N) - j\, sin(2\pi nm/N)], \quad (4)$$

onde

m = índice da saída da DFT no domínio da frequência;

n = índice das amostras de entrada no domínio do tempo;

X(m) = m-éseima saída da DFT;

x(n) = n-ésima entrada da DFT;

N = número de dados usados como entrada da DFT e o número de pontos de frequência de saída da DFT.

O grande problema com a aplicação direta da DFT é a sua ineficiência. Na medida que cresce o número de pontos a quantidade de operações matemáticas necessárias torna-se excessivamente alto. Em 1965 Cololey e Tukey publicaram um artigo descrevendo um algoritmo eficiente para a implementação da DFT que passou a ser conhecido como transformada rápida de Fourier (FFT). Para uma DFT com *N* pontos amostrados, é necessário um total de N^2 multiplicações complexas. Esse número cai para $N/2 \log_2 N$ com o uso da FFT. Lyons (2011) em seu livro cita um exemplo que mostra esse aumento drástico de eficiência, caso a operação de uma FFT de dois milhões de pontos (*N*=2.097.152) levasse 10 segundos em um computador, com o mesmo equipamento a DFT levaria mais de três semanas.

3.3. Análise de Componentes Principais

A análise de componentes principais (PCA) é uma técnica empregada para reduzir o número de dimensões presentes em um conjunto de dados. As dimensões são as

variáveis, chamadas do contexto do aprendizado de máquina de atributos. Neste trabalho cada atributo é uma frequência do espectro do sinal acústico captado pelos microfones.

Geralmente quando o conjunto de dados é grande, parte das variáveis costumam estar correlacionadas entre si. O PCA então transforma esse conjunto grande de variáveis em outro menor, os componentes principais, não correlacionados entre si e que contêm a maior parte da informação presente nos dados originais. Nesse processo as variáveis correlacionadas não são simplesmente eliminadas, os componentes principais criados são combinações lineares das variáveis originais. Para ilustrar a drástica redução no número de variáveis possível de ser alcançada com essa técnica, FERNANDES et al. (2016) aplicaram o PCA ao espectro do sinal acústico capturado por microfones, o que resultou na redução de 2305 amplitudes no domínio da frequência a 200 componentes principais, os quais foram suficientes para representar mais de 95% da variância presente nos dados originais. Descrição detalhada e rigorosa do PCA pode ser encontrada em JOLLIFFE (2002).

3.4. Algoritmos de Aprendizado de Máquina

O aprendizado de máquina é o campo da ciência que estuda algoritmos e métodos estatísticos que são desenvolvidos para aprender a partir de dados que lhes são fornecidos. Esses métodos são capazes de automaticamente identificar padrões presentes nos dados, sem que lhes seja passado nenhum conjunto específico de instruções. Esse conhecimento adquirido passa então a guiar a tomada de decisões dos modelos.

Os algoritmos de aprendizado de máquina empregados neste trabalho foram: K Vizinhos Mais Próximos (KNN), Regressão Logística (LR), Redes Neurais (NN), Máquinas de Vetores Suporte com kernel linear (SVM-LINEAR), Máquinas de Vetores Suporte com kernel função de base radial (SVM - RBF), Floresta Aleatória (RF), *Adaptive Boosting* (Adaboost), e *Extreme Grandient Boosting* (Xgboost). Esses algoritmos vêm sendo aplicados com sucesso em diversos campos de pesquisa. KNN é um algoritmo muito conhecido, que a despeito da sua simplicidade, pode funcionar bem em diversas situações (JAMES et al., 2013). A técnica regressão logística foi aplicada com sucesso na detecção de problemas em válvulas de compressores. O modelo identificou as falhas no equipamento a partir de padrões encontrados em dados de vibração obtidos a partir do monitoramento do sistema com o uso de acelerômetros (PICHLER et al., 2016). Máquinas de Vetores Suporte foram empregadas para a identificação de vazamentos em

tubulações transportando água (KAYAALP et al., 2017) e gás (JIN et al., 2014).SANTOS et al. (2014) aplicaram redes neurais para a detecção, localização e determinação da magnitude de vazamentos em um tubulação de gás. Os modelos Floresta Aleatória e Adaboost não são ideias novas, foram propostos em 1995 e 1998 respectivamente, porém se mantem como duas das mais poderosas ferramentas de aprendizado de máquina disponíveis (GÉRON, 2017).CHO et al. (2018) propuseram o uso dos modelos Floresta Aleatória e Redes Neurais Artificiais para a identificação de vazamentos em indústrias químicas. Os autores reportaram performance superior do algoritmo Floresta Aleatória. Já o modelo Xgboost foi empregado em novo método proposto para detecção de falhas em turbinas eólicas, tendo sucesso quando aplicado a vários modelos de turbinas operando nas mais diversas condições (ZHANG et al., 2018).

Os próximos subtópicos contêm breves explicações a respeito do funcionamento dos algoritmos empregados. Um tratamento introdutório e voltado a aplicações desses algoritmos (exceção a redes neurais) pode sem encontrado em JAMES et al. (2013), enquanto tratamento mais matematicamente rigoroso pode ser visto em HASTIE; TIBSHIRANI; FRIEDMAN (2017). Já informações a respeito de Redes Neurais Artificiais podem ser encontradas em HAYKIN (2009).

3.4.1. K Vizinhos Mais Próximos

K Vizinhos Mais próximos é um algoritmo simples que pode ser usado tanto para classificação quanto para regressão, mas que é principalmente empregado em problemas de classificação. Essa técnica se baseia na ideia de que coisas semelhantes tendem a estar próximas umas das outras, e que cada uma delas pode ser representada com um ponto no espaço. Quando o modelo recebe um elemento qualquer x que se deseja classificar, o algoritmo calcula a distância entre x e cada uma das amostras que compõe uma dada base de dados, onde a classe de cada um dos elementos dessa base já é conhecida. A x é então atribuída a classe da maioria dos elementos dentre os k que estão mais próximos a ele, daí o nome K Vizinhos Mais Próximos. O valor de k deve ser definido pelo usuário (é um hiperparâmetro) e sua escolha tem papel fundamental nos resultados obtidos pelo modelo. O algoritmo funciona da mesma maneira para regressões; a diferença está na reposta do modelo, que se torna a média dos valores dos k vizinhos mais próximos.

K Vizinhos Mais Próximos pertence a uma classe de algoritmos chamados de preguiçosos. Algoritmos de aprendizado de máquina costumam aprender uma função

geral a partir de um conjunto de exemplos fornecidos aos modelos. A partir daí essa função é usada classificar outros elementos. Já algoritmos preguiçosos somente armazenam os dados de treinamento, cada nova instância é classificada com base na similaridade com as instâncias presentes nos dados de treinamento.

3.4.2. Regressão Logística

Regressão logística é um algoritmo de classificação semelhante a regressão linear, mas ao contrário desta última, cujo modelo fornece como resposta um valor numérico contínuo para uma variável, a saída da regressão logística é a probabilidade que um certo elemento pertença a uma classe. Além disso, ao invés de tentar ajustar uma reta ou superfície ao conjunto de dados como faz a regressão linear, a regressão logística usa um tipo de função que mantém a resposta do modelo restrita ao intervalo [0,1]. Como o nome da técnica sugere, a função usada é logística:

$$p(X) = \frac{e^{\beta_0 + \beta_1 X_1 + \ldots + \beta_p X_p}}{1 + e^{\beta_0 + \beta_1 X_1 + \ldots + \beta_p X_p}} \quad (5)$$

Na Equação (5) $X = (X_1, \ldots, X_p)$ é um conjunto de preditores, p é a probabilidade que o modelo fornece como resposta e $\beta_0 \ldots \beta_p$ são os parâmetros do modelo que são ajustados durante a etapa de treinamento.

O pressuposto desta técnica é que elementos de duas classes que se desejam classificar podem ser separados em duas regiões distintas, e a superfície que divide essas duas regiões é linear (Figura 7). O modelo foi desenvolvido para lidar com a classificação de elementos de duas classes distintas, mas existem diversas estratégias disponíveis que o tornam apto a lidar com problemas que envolvem mais de duas classes.

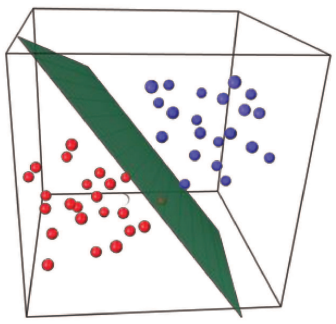

Figura 7 - Separação elementos de duas classes por superfície linear. FONTE: Joglekar (2015)

3.4.3. Máquinas de Vetores Suporte

Máquinas de vetores suporte é uma técnica que pode ser empregada tanto para classificação quanto para regressão. Como técnica de classificação, foi desenvolvida para lidar com somente duas classes, mas da mesma forma que a regressão logística, pode ser empregada para classificar elementos em mais de duas classes. Seu precursor foi o método de classificação simples chamado classificador de margem máxima. Esse classificador se aplica somente a dados que são linearmente separáveis (Figura 7). O algoritmo busca um hiperplano (por exemplo, em duas e três dimensões esse hiperplano é uma reta e um plano respectivamente) que divida os dados em duas regiões, de modo que dados da mesma classe estejam de um mesmo lado do hiperplano. Dessa forma os dados são classificados de acordo com o lado que ocupam em relação ao hiperplano. Mas, casos os dados sejam linearmente separáveis, existem infinitos hiperplanos que podem separá-los. Assim, dentre todos eles, o algoritmo busca o hiperplano cuja distância (essa distância é chamada de margem) em relação aos elementos das classes opostas é máxima. (Figura 8). Espera-se que este hiperplano apresente o melhor desempenho na tarefa de classificar dados que não foram usados na construção do modelo, ou seja, dados que não foram usados no treinamento.

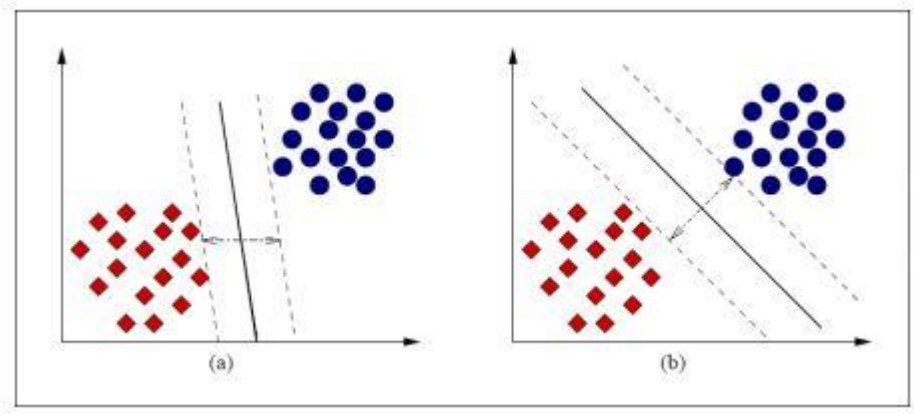

Figura 8 - Separação de dados de duas classes: pontos azuis e vermelhos a) Hiperplano de separação pequena. b) Hiperplano de separação máxima. Fonte: Meloni (2019).

Infelizmente a maioria dos problemas que envolvem classificação é composta de dados que não são linearmente separáveis, casos nos quais o hiperplano de distância máxima não existe. Para lidar com os casos nos quais esse hiperplano não existe, mas nos quais os dados podem ser satisfatoriamente separados por uma superfície linear, foi criada uma generalização do classificador de margem máxima chamado de classificador de vetores suporte. O classificador de vetores suporte não busca o hiperplano que classifique

corretamente a totalidade dos dados de treinamento, ao invés disso, ele busca o hiperplano que classifique corretamente a maioria dos elementos das duas classes e que ao mesmo tempo esteja a máxima distância da maioria dos elementos das duas classes opostas, ou seja, na busca do hiperplano o modelo permite que alguns elementos não respeitem a máxima distância e até mesmo sejam classificados incorretamente. Essa abordagem torna o modelo apto a lidar com dados que não são linearmente separáveis, além de aumentar sua capacidade de generalização e de lidar com valores atípicos ("*outliers*"). A quantidade de elementos que poderão não respeitar a distância máxima é definida pelo usuário. Uma menor tolerância a violações corresponde a um melhor ajuste aos dados de treinamento e consequentemente a uma margem mais estreita, o que pode resultar em resultados ruins quando o modelo for testado com dados não usados no treinamento, fato conhecido como "*overfitting*". Enquanto uma maior tolerância leva a margens maiores (Figura 9). Uma característica importante dessa técnica é que a mudança de posição de elementos que estão além da distância máxima, desde que não passem a violar a margem, não altera o hiperplano. Somente os elementos que estão na margem e os que a violam, corretamente e incorretamente classificados, influenciam na escolha do hiperplano. Esses elementos são chamados de vetores suporte, daí o nome do modelo.

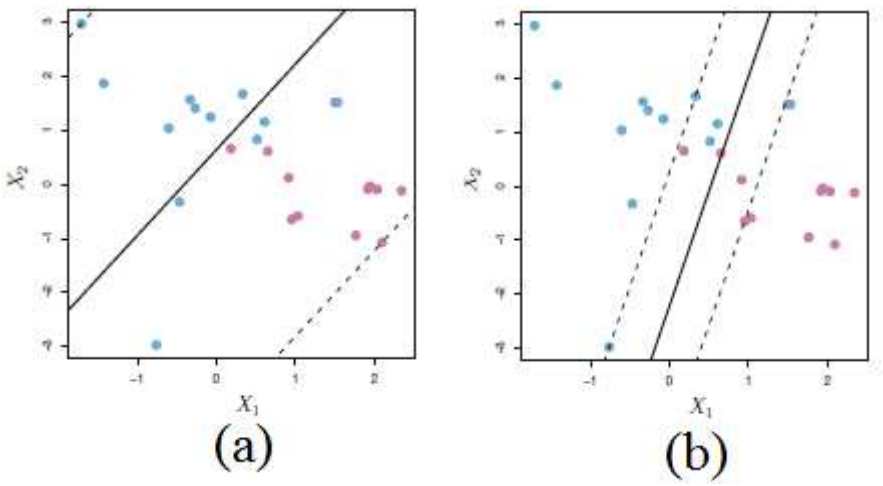

Figura 9 - a) Classificador de vetores suporte com grande tolerância a violações da margem, e consequentemente com uma margem larga. b) pequena tolerância a violações, o que leva a uma margem estreita. Fonte: JAMES et al., (2013)

. Os Classificadores de Vetores Suporte são eficientes, porém sua aplicabilidade é limitada. Nem sempre é possível, na verdade na maioria dos casos não é, dividir de maneira adequada os dados com o uso de uma superfície de separação linear (Figura 10). Para lidar com esse tipo de dados, uma generalização dos classificadores de vetores suporte pode ser aplicada, conhecida como Máquinas de Vetores Suporte (SVM)

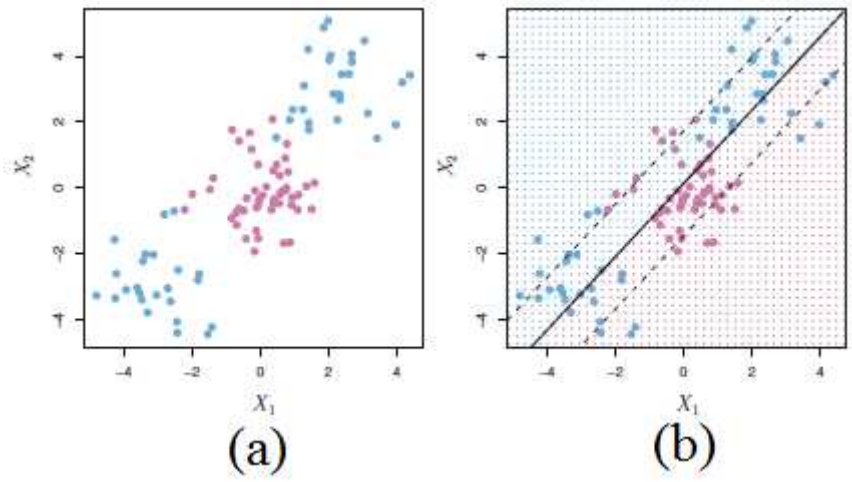

Figura 10 - (a). Dados com duas classes que não podem ser separados com o uso de uma superfície linear. (b) Tentativa de classificação dos dados com o uso do classificador de vetores suporte. Como os dados não podem ser separados por uma superfície linear, a performance do modelo é ruim. Fonte: JAMES et al., (2013).

As Máquinas de Vetores Suporte buscam mapear os dados de entrada do modelo do espaço original para um novo espaço de maiores dimensões no qual os dados sejam linearmente separáveis (Figura 11). Essa mudança é conduzida através da aplicação de uma técnica conhecida como truque kernel. A tarefa de encontrar o melhor hiperplano que divide os dados de duas classes é um problema de otimização que só envolve o cálculo dos produtos escalares entre os dados de treinamento (HASTIE; TIBSHIRANI; FRIEDMAN, 2017). Quando se usa o espaço com maiores dimensões, a escolha do hiperplano passa a envolver os produtos escalares dos dados nesse novo espaço. O truque kernel permite o cálculo desses produtos escalares no novo espaço sem que seja necessário o cálculo das coordenadas das variáveis no novo espaço, nem mesmo é necessário conhecer a função que transforma os dados originais para o novo espaço. Sem a aplicação do truque kernel seria necessário obter as coordenadas no novo espaço e depois calcular os produtos escalares. Com o kernel os produtos escalares são calculados diretamente. Ganha-se assim eficiência computacional.

Tratou-se até aqui de máquinas de vetores suporte aplicados a classificação. O modelo funciona de forma semelhante quando aplicado a regressões. Nesse caso, o objetivo do algoritmo é oposto, ao invés de buscar o hiperplano que melhor separe os dados buscando limitar as violações as margens, busca-se um hiperplano cujas margens contenham o maior número possível das amostras de treinamento, buscando-se limitar o número de amostras que estejam fora das margens (Figura 12).

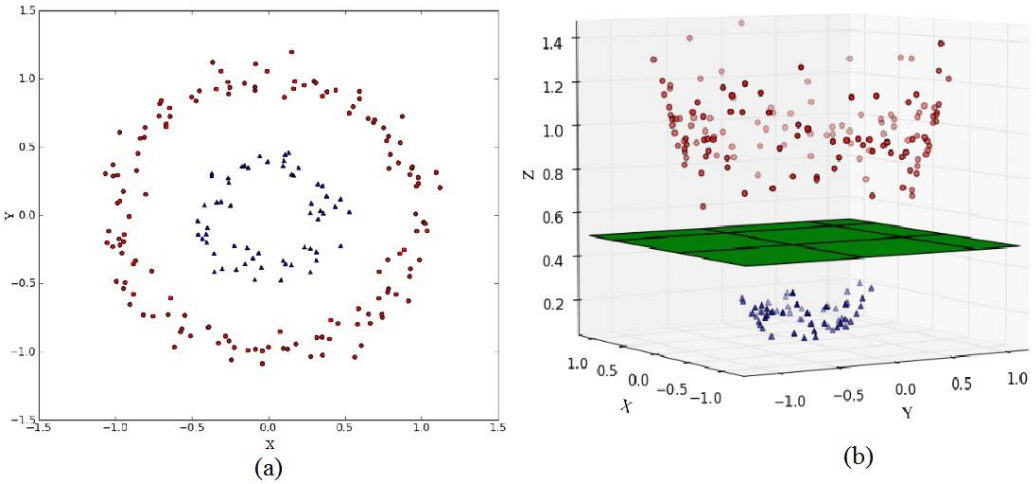

Figura 11 - (a) Dados não linearmente separáveis por um hiperplano no espaço de duas dimensões R^2. (b) Dados tornam-se linearmente separáveis ao serem transportados de R^2 para R^3. Fonte: Haltuf (2014).

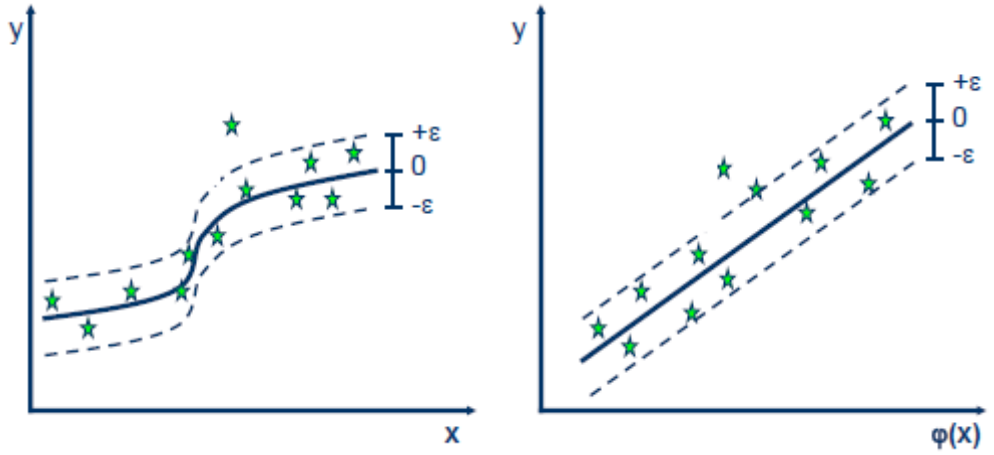

Figura 12 - - Duas aplicações de SVM a problemas de regressão. Busca-se as margens (linhas pontilhadas) que contenham o maior número das amostras dos dados de treinamento. Ao mesmo tempo, em busca do melhor ajuste, permite-se que algumas amostras a violem. Fonte: Sayad (2017).

3.4.4. Redes Neurais Artificiais

Redes Neurais Artificiais são modelos computacionais inspirados no sistema nervoso dos seres vivos. São processadores paralelos maciçamente distribuídos formados por unidades de processamentos simples, os neurônios artificiais. Essa tecnologia pode ser aplicada no reconhecimento e classificação de padrões, no agrupamento de dados, em sistemas de previsão, na otimização de sistemas, sendo especialmente apropriadas para sistemas não lineares, que dão origem a problemas de difícil tratamento matemático (NUNES; SPATTI; FLAUZINO, 2016).

Os neurônios artificiais são as unidades básicas que formam uma rede neural. Essas unidades são conectadas por canais de comunicação associados a determinado peso e que

fazem operações sobre seus dados locais, que são as entradas recebidas pelas suas conexões. A operação de um neurônio artificial pode ser resumida da seguinte forma:

- Os sinais de entrada são recebidos;
- Cada sinal de entrada é multiplicado por um peso, que representa a sua influência na saída que será gerada;
- É feita a soma ponderada dos sinais que produz um nível de atividade;
- Se este nível de atividade exceder um certo limite (*threshold*) a unidade produz uma determinada resposta de saída.

O modelo de um neurônio artificial é apresentado na Figura 13. Nela aparecem as entradas fornecidas, a saída gerada, os pesos sinápticos (w_k) pelo quais as entradas são multiplicadas, a função de ativação que controla a saída do neurônio, e um bias, representado por b_k, que são aplicadas para aumentar ou reduzir a entrada na função de ativação.

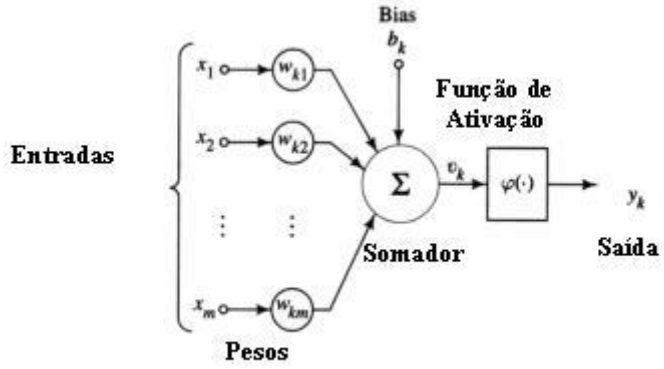

Figura 13 - Neurônio artificial. Fonte: Adaptado de HAYKIN; ENGEL (2001).

As redes neurais são divididas em camadas, classificadas como camada de entrada, camadas intermediárias ou escondidas e camada de saída. Na primeira os dados de entrada são apresentados à rede e associados a um peso de entrada, na segunda são extraídas as informações do sistema estudado e é onde a maior parte do processamento ocorre e por fim na última os dados das camadas anteriores são agregados e é gerada uma resposta (NUNES; SPATTI; FLAUZINO, 2016).

A arquitetura e a topologia são dois fatores de distinção entre redes neurais. A arquitetura refere-se ao modo como os neurônios estão agrupados e a topologia é a forma como os neurônios se conectam para formar a rede. Outra distinção diz respeito ao tipo de aprendizado, que pode ser supervisionada e não-supervisionada. No treinamento

supervisionado são necessárias as amostras de entrada e as respectivas saídas desejadas para que o algoritmo ajuste os pesos, ou seja, o modelo matemático é ajustado com o conhecimento tanto de seu ponto de partida (as entradas) como do seu ponto de chegada (as saídas). A rede assim treinada gera uma saída numérica. Já no treinamento não supervisionado somente as entradas são fornecidas, a saída deve se auto organizar de acordo com as características das entradas e assim identificar subconjuntos de acordo com alguma similaridade (HAYKIN; ENGEL, 2001).

3.4.5. Árvores de Decisão

Árvores de Decisão é um algoritmo de aprendizado de máquina que pode ser usado para problemas de classificação e de regressão. Ele pode ser aplicado satisfatoriamente a muitos conjuntos de dados. Além disso, é um componente básico de alguns dos algoritmos de aprendizado de máquina mais poderosos disponíveis atualmente, como Florestas Aleatória, Xgboost e Adaboost (o componente básico do Adaboost não necessariamente é uma Árvore de Decisão) (JAMES et al., 2013).

O modelo Árvores de Decisão prediz uma classe ou um valor através da construção de um fluxograma. Cada nó do fluxograma é uma pergunta referente a um dos atributos presentes nos dados apresentados ao modelo. O resultado de cada pergunta é a divisão dos dados, que geram os ramos do fluxograma. O atributo em cada nó é selecionado de forma a produzir a melhor divisão dos dados, ou seja, busca-se a maior semelhança possível entre os dados em cada ramo. Os nós finais são chamados de folhas, novos elementos são classificados de acordo com a folha em que sejam posicionados. Ela atribui ao novo elemento a mesma classe presente na maioria dos elementos dos dados de treinamento presentes nessa folha. No caso de regressão é atribuída a média dos valores dos elementos na folha. A Figura 14 mostra uma Árvore de Decisão aplicada a classificação de flores, que podem pertencer a três classes: setosa, versicolor e virginica. O primeiro bloco do fluxograma mostra que o conjunto de dados inicial continha 150 amostras, divididos igualmente entre as três classes; *elementos de cada classe* (variável presente nos blocos do fluxograma*)*, como o nome sugere, indica o número de elementos de cada classe, o primeiro número nos colchetes refere-se a classe setosa, o segundo versicolor e o terceiro a virginica ; *gini* é uma métrica que indica a pureza dos dados, que pode ser calculada pela Equação (6), onde $p_{i,k}$ é a probabilidade de que um elemento em um nó i pertença a cada uma das classes k. Por exemplo, no fluxograma o atributo empregado no primeiro nó foi o comprimento das pétalas, que gerou em um dos ramos

dados completamente puros, todos os elementos pertencentes a uma mesma classe, o que corresponde a *gini* nula. Assim, todas as amostram apresentadas a essa Árvore de Decisão que possuam pétalas com comprimentos menores ou iguais a 2,45 cm serão classificadas como setosa.

$$G_i = 1 - \sum_{k=1}^{n} p_{i,k}^{2} \tag{6}$$

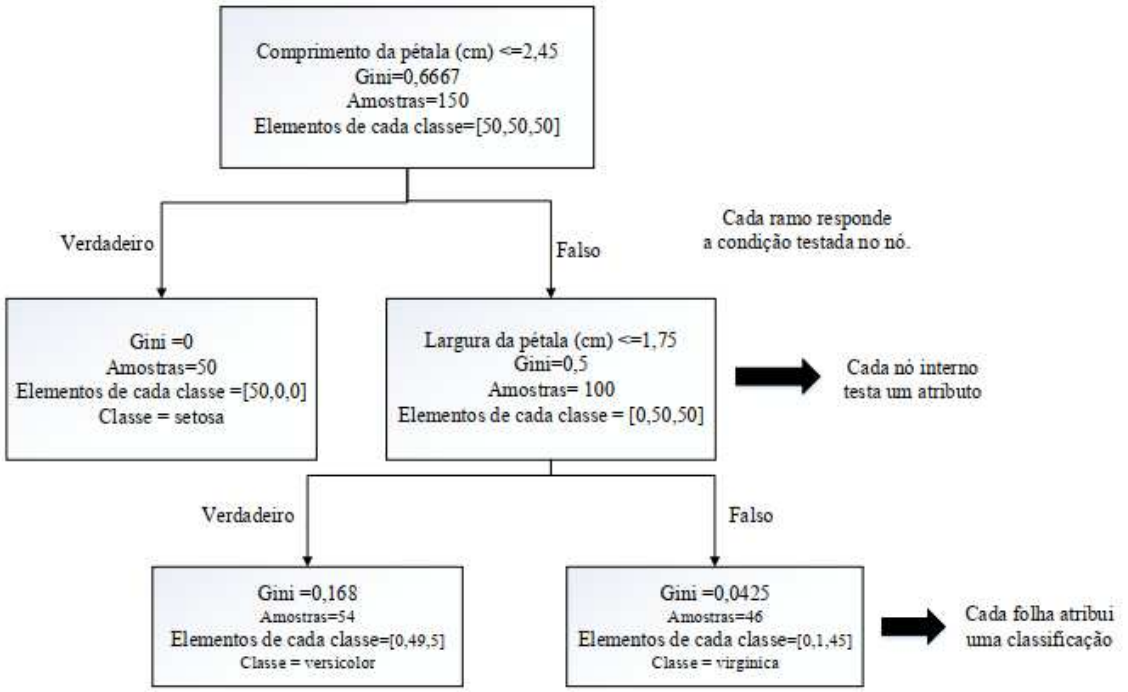

Figura 14 - Aplicação de Árvore de Decisão para a classificação de flores. Fonte: Adaptado de GÉRON (2017).

Árvores de Decisão é um algoritmo simples, intuitivo e poderoso, porém para gerar bons resultados geralmente são necessárias árvores grandes (muitos nós), o que as tornam suscetíveis a *overfitting,* situação na qual o modelo se ajusta demasiadamente aos dados de treinamento, obtendo assim baixa performance quando são apresentados novos dados. Essa limitação pode ser enfrentada com o uso de um conjunto de técnicas conhecidas como *ensemble* ou sistemas de classificadores múltiplos. A discussão aqui será restrita ao uso de Árvores de Decisão como preditor base, mas os sistemas de classificadores múltiplos não se restringem a elas. A ideia dessas técnicas é combinação de diversas preditores para gerar resultados superiores aos que seriam obtidos com os preditores quando empregados individualmente. As Árvores de Decisão que compõem o conjunto de classificadores têm seu tamanho limitado, evitando-se assim o *overfitting*. O resultado do modelo é alguma forma de combinação dos resultados obtidos por cada um

dos preditores. Os sistemas de classificadores múltiplos podem ser de dois tipos: paralelos, Floresta aleatória é um exemplo e sequenciais, como Adaboost e XgBoost.

O modelo Floresta aleatória monta *k* subconjuntos de dados, cada um deles contendo o mesmo número de elementos. A seleção dos componentes de cada subconjunto é feita com reposição, assim os subconjuntos se sobrepõem de forma significativa e os dados podem aparecer de forma repetida até dentro de um mesmo subconjunto. Os dados são usados para treinar *k* (*k* é um hiperparâmetro definido pelo usuário) Árvores de Decisão, cada uma delas com os elementos de um dos *k* subconjuntos. Para aumentar a variabilidade entre os modelos, a cada um deles é permitido selecionar os atributos para cada nó dentre um conjunto restrito selecionado aleatoriamente dentre todos os disponíveis. Busca-se a assim maior variabilidade entre os modelos, de modo que os mesmos tipos de erros não sejam cometidos por diferentes árvores. A predição final de uma Floresta Aleatória será o resultado predito pela maioria das *k* Árvores de Decisão.

O modelo Floresta Aleatória, além do seu uso em problemas de classificação e regressão, possibilita medir a importância de cada um dos atributos presentes nos dados. Ela é medida pela redução na impureza dos dados que resulta do uso de um certo atributo em um nó de uma árvore de decisão. Quanto maior a redução, mais importante o atributo. Além disso, leva-se também em conta no cálculo o número de amostras presentes nos nós, quanto maior o número de amostras um peso maior é atribuído ao atributo. Esse método permite então a seleção somente dos atributos que tenham maior influência nas decisões dos modelos, descartando-se os dados não relevantes e consequentemente facilitando o tratamento dos dados.

Os sistemas de classificadores múltiplos sequenciais funcionam de forma diferente. Neles os preditores são treinados sequencialmente, cada um deles tentando corrigir os erros cometidos pelos preditores treinados anteriormente. O treinamento prossegue até que se atinja o número de preditores máximo definido pelo usuário ou se atinja alguma condição estabelecida.

O modelo Adaboost inicialmente atribui um mesmo peso a todos os elementos dos dados de treinamento. Em seguida um primeiro preditor é construído com um subconjunto aleatório dos dados. Após o treinamento os elementos corretamente classificados têm seus pesos reduzidos e os incorretamente classificados tem seus pesos incrementados. Esses pesos são usados na escolha dos elementos que serão usados no treinamento do preditor seguinte. Os elementos são selecionados com reposição dentre o

conjunto de dados, quanto maior o peso de um elemento maior a probabilidade de ele ser escolhido. Com a escolha é feita com reposição, um mesmo elemento pode aparecer diversas vezes nos dados de treinamento de um mesmo preditor. Com essa forma de seleção, cada novo preditor será direcionado as amostras erroneamente classificadas anteriormente. Não somente os elementos recebem um peso, aos preditores também são atribuídos pesos, quanto melhor a performance maior é o peso. Assim o resultado do modelo é a combinação dos resultados de cada um dos preditores, com os preditores com maiores pesos tendo maior influência na resposta final.

O modelo Xgboost também é um classificador múltiplo sequencial, mas funciona de maneira ligeiramente diferente do método Adaboost. A cada nova árvore adicionada o modelo Xgboost tenta ajustar o novo preditor aos resíduos do modelo anterior. Os resíduos são a diferença entre os valores reais e os preditos pelo modelo. Assim, cada novo preditor, como no caso do modelo Adaboost mas por um meio diferente, concentra-se nos erros do anterior. A primeira Árvore de Decisão é ajustada aos dados, com base nas predições desse primeiro modelo são calculados os primeiros resíduos. O segundo preditor tem como entrada os mesmos dados do preditor anterior, mas como saída os resíduos calculados, ou seja, a cada dado de entrada é atribuído um resíduo. Após o ajuste do modelo os resíduos são usados para atualizar as predições anteriores, e essas predições atualizadas são usadas para calcular novos resíduos. O processo prossegue até que os valores dos resíduos se tornem estáveis ou o número máximo de preditores estabelecido seja atingido.

4.Materiais e Métodos

Os experimentos deste trabalho foram realizados em aparato experimental montado no Laboratório de Controle e Automação de Processos (LCAP) da Faculdade de Engenharia Química (FEQ) da Unicamp. As ondas sonoras geradas durante a operação da tubulação, em condições normais e na presença de vazamentos, foram captadas por microfones. Algoritmos de aprendizado de máquina implementados tanto para identificar quanto para localizar os vazamentos foram alimentados com o sinal captado pelos microfones, buscando-se assim um sistema eficiente de monitoramento de vazamentos em uma tubulação de gás operando sob baixa pressão. Inicialmente o aparato experimental será descrito e em seguida a metodologia empregada.

4.1. Aparato Experimental

O equipamento experimental é composto de uma válvula reguladora de pressão, um vaso de armazenamento, tubulação de cobre, sensores para captura do sinal acústico (microfones) e o sistema de aquisição de dados. Como fluido de trabalho foi usado o ar comprimido proveniente da linha de alimentação da faculdade.

O regulador de pressão foi usado para ajustar a pressão no início de cada experimento. A pressão manométrica empregada foi de 1 kgf cm^{-2}, baseada na pressão nas linhas de distribuição de gás doméstico. O vaso de armazenamento é um botijão de cozinha convencional de aço carbono com capacidade de 34,5 L. A tubulação de cobre tem aproximadamente 53 m de comprimento e ½'' de diâmetro (Figura 15). Nela foram perfurados, com o uso de uma broca, sete orifícios em diferentes posições e de diferentes diâmetros, conforme especificado na Figura 16. Esses furos permaneceram cobertos por fita isolante para evitar a fuga de ar comprimido.

Figura 15 - Tubulação montada no LCAP da FEQ/Unicamp.

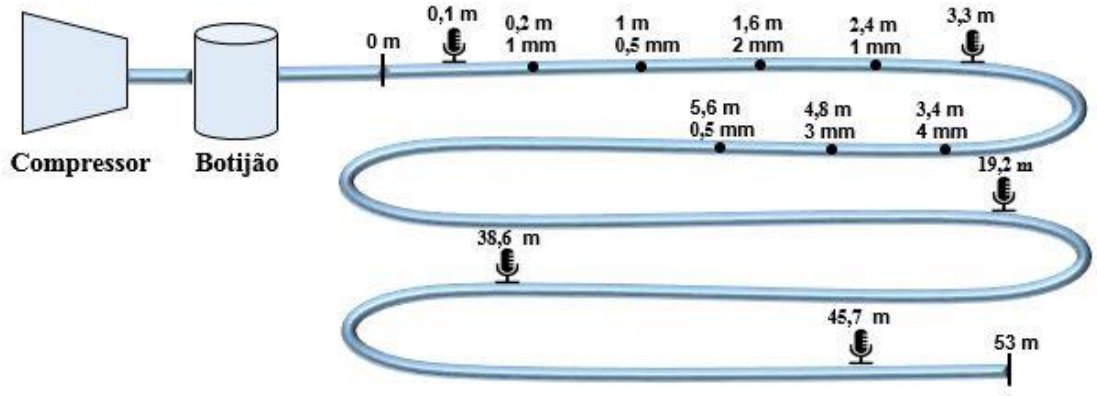

Figura 16 - Posições e diâmetros dos orifícios e posição dos microfones.

4.1.1. Sensores

Microfones foram empregados para a captura do sinal acústico. Esse tipo de sensor converte ondas sonoras em um sinal elétrico. Os microfones utilizados são do modelo XCM 9767, omnidirecionais, ou seja, capazes de captar sons provenientes de todas as direções. A faixa de operação desses instrumentos vai de 50 Hz até aproximadamente 16 kHz (Figura 17).

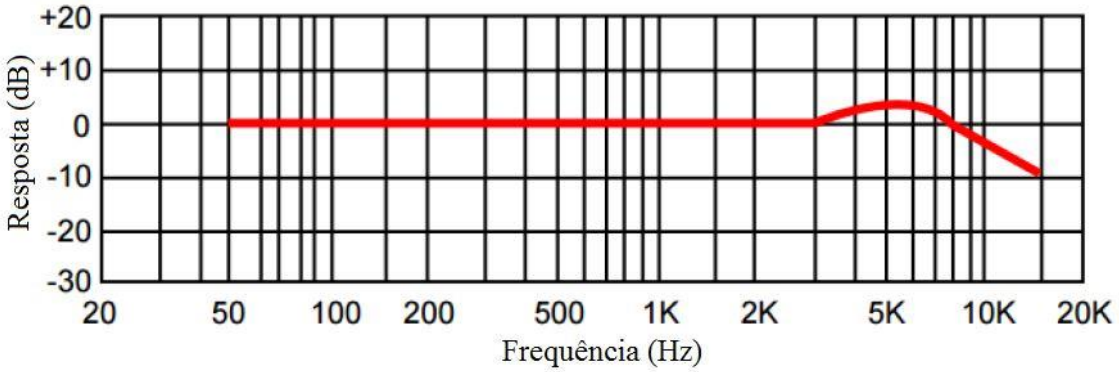

Figura 17 - Resposta em frequência dos microfones modelo XCM 9767. Fonte: Santos (2015).

Quinze microfones foram instalados ao longo da tubulação, mas dada a limitação da placa coletora de dados apenas cinco deles foram utilizados no presente trabalho (Tabela 2).

Tabela 2 - Posição dos microfones.

Microfone	Distância (m) em relação a entrada da tubulação.
1	0,1
2	3,3
3	19,2
4	38,6
5	45,7

O sinal captado pelos microfones passou por pré-amplificadores, e em seguida o sinal analógico foi convertido em sinal digital pela placa conversora A/D NI cDAQ-9178. Por fim o sinal digital foi então enviado a um microcomputador com o *software Labview*.

4.2. Métodos

Serão descritos nessa seção os procedimentos que foram executados neste trabalho. Começando pela descrição dos parâmetros adotados para a coleta dos dados, seguindo pelos experimentos realizados e o tratamento dos dados coletados e por fim a implementação dos algoritmos de aprendizado de máquina responsáveis por identificar e localizar os vazamentos na tubulação.

4.2.1. Aquisição dos dados

A rotina para aquisição do sinal acústico captado pelos microfones foi implementada no *software Labview*. No ambiente de aquisição de dados foi adicionado um filtro virtual, cuja finalidade foi atenuar frequências indesejáveis, abrandando dessa forma uma fração do sinal que não teria utilidade para a finalidade pretendida neste trabalho. O filtro ativo *Butterworth* de 10º ordem foi usado, com base no trabalho de Santos (2015). Esse é um filtro passa faixa que combina um filtro passa baixa e um filtro alta, assim frequências acima e abaixo de determinados valores estipulados foram atenuadas. Os valores adotados foram de 100 Hz e 10.000 Hz. A escolha desses dois valores seguiu trabalho de Santos (2015), que mostrou que a maior parte da energia do sinal acústico gerado pela operação da tubulação concentra-se neste intervalo.

A frequência de amostragem empregada foi de 30.000 Hz. Essa frequência é maior que o dobro da maior frequência em análise (10.000 Hz). Isso garantiu, de acordo com o teorema de *Nyquist*, a não ocorrência do fenômeno de *aliasing*.

4.2.2. Experimentos

Quatorze tipos de experimentos foram executados, divididos em três categorias: experimentos sem vazamento, experimentos com um único vazamento e experimentos com dois vazamentos simultâneos (Tabela 3).

Em dois dos três experimentos sem vazamentos foram provocadas perturbações no sistema: batidas na tubulação e batidas no vaso de armazenamento. Perturbações como essa são comuns na operação de tubulações. Elas representam um sério desafio para sistemas automáticos de detecção de vazamentos, uma vez que as perturbações tendem a

gerar falsos alarmes. Assim os experimentos com perturbações testaram a capacidade do sistema proposto de distinguir vazamentos de outros tipos de perturbações no sistema.

Tabela 3 - Experimentos realizados.

Experimento	Descrição
1	Sem Vazamentos
2	Sem Vazamentos. Batidas no botijão.
3	Sem vazamentos. Batidas na tubulação.
4	Vazamento no orifício 1
5	Vazamento no orifício 2
6	Vazamento no orifício 3
7	Vazamento no orifício 4
8	Vazamento no orifício 5
9	Vazamento no orifício 6
10	Vazamento no orifício 7
11	Vazamentos nos orifícios 1 e 3
12	Vazamentos nos orifícios 1 e 4
13	Vazamentos nos orifícios 1 e 7
14	Vazamentos nos orifícios 2 e 7

O primeiro passo no procedimento experimental foi o ajuste da pressão para 1 kgf cm^{-2}. Durante todo o experimento a válvula de alimentação de ar comprimido foi mantida aberta. A duração de cada experimento foi de 60 s. Os vazamentos foram provocados manualmente, e tanto os vazamentos quanto as perturbações externas iniciaram-se 10 s após o início de cada experimento.

Ao final de cada experimento foram coletadas 1.800.000 (frequência de amostragem multiplicada pelo intervalo de tempo dos experimentos) amplitudes no domínio do tempo para cada um dos cinco microfones.

4.2.3. Processamento dos dados

O processamento de dados foi realizado utilizando-se bibliotecas de código aberto da linguagem de programação *Python*. A primeira etapa consistiu na conversão dos dados coletados pelos microfones do domínio do tempo para o domínio da frequência. Para esse fim foi usada a biblioteca de computação científica *SciPy* (OLIPHANT, 2007). Na segunda etapa duas técnicas de redução de dimensionalidade foram empregadas para facilitar o trabalho dos algoritmos de aprendizado de máquina. Essas técnicas foram aplicadas com o uso da biblioteca de algoritmos de aprendizado de máquina *Scikit-learn* (PEDREGOSA et al., 2011). Essas duas etapas são descritas com mais detalhes a seguir.

A FFT foi aplicada para a conversão dos dados do domínio do tempo para o domínio da frequência. Só foram usadas as amplitudes coletadas após o sistema atingir o estado estacionário, que foi alcançado aproximadamente 10 s após o início dos vazamentos. Dessa forma, como os vazamentos foram provocados 10 s após o início dos experimentos, somente foram usadas as amplitudes coletadas 20 s após o início destes. Assim 600.000 amplitudes das 1.800.000 de cada um dos experimentos foram descartadas.

As 1.200.000 amplitudes no domínio do tempo foram fatiadas em grupos de 10.000. Cada grupo de 10.000 amplitudes consistiu em uma amostra dos experimentos, essas amostras serão posteriormente apresentadas aos algoritmos de aprendizado de máquina. A cada uma dessas amostras foi aplicada a FFT. Assim 10.000 amplitudes no domínio do tempo foram transformadas em 10.000 amplitudes no domínio da frequência. Dada a simetria da FFT, 5.000 de cada uma das 10.000 amplitudes foram descartadas. Além disso, foram descartadas também as frequências atenuadas pelo filtro (inferiores a 100 Hz e superiores a 10.000 Hz). Descartadas as frequências atenuadas as 5.000 amplitudes foram reduzidas a 3340. Ao final, após a aplicação da transformada, os dados de cada experimento foram convertidos em 120 amostras (1.200.000 amplitudes agrupados em grupos de 10.000), com cada uma das amostras contendo 3340 atributos (amplitudes no domínio da frequência.).

Lidar com dados contendo 3340 dimensões seria muito custoso computacionalmente. Mas não somente isso, o número elevado de dimensões poderia reduzir a performance dos algoritmos de aprendizado de máquina (JAMES et al., 2013).O primeiro método usado para reduzir o número de dimensões foi a análise de componentes principais (PCA), técnica aplicada com sucesso por FERNANDES; SANTOS; FILETI (2016) ao lidar com o mesmo tipo de sinal. As 3340 dimensões foram reduzidas a 287, 236, 102, 77 e 81 componentes principais nos dados coletados pelos microfones 1 a 5. Este número de componentes continha 99% de toda a variância presente nos dados originais. Já o segundo método empregou o algoritmo de aprendizado de máquina Floresta Aleatória conforme proposto por ZHANG et al. (2018). O método selecionou dentre os 3340 atributos somente aqueles considerados mais importantes. A importância de um atributo foi medida pela queda de impureza nos dados que resultou de seu uso em um nó de uma árvore de decisão. Somente os atributos cujas importâncias eram superiores à média das importâncias de todas os atributos foram mantidas. O número de atributos considerados mais importantes encontrados foi de 225, 347, 361, 258 e 154 nos dados

dos microfones de 1 a 5. Concluída a redução de dimensionalidade, os dados foram empregados nos algoritmos de aprendizado de máquina.

4.2.4. Implementação dos Algoritmos de Aprendizado de Máquina

Os algoritmos de aprendizado de máquina empregados foram descritos na seção 'Algoritmos de Aprendizado de Máquina'. Todos eles foram aplicados com o uso da biblioteca *Scikit-learn* (PEDREGOSA et al., 2011).

As técnicas de aprendizado de máquina foram aplicadas com duas finalidades distintas: classificação e regressão. A tarefa dos algoritmos de classificação consistiu em identificar a qual dos quatorze experimentos realizados as entradas apresentadas aos modelos pertenciam. Dessa forma os algoritmos foram alimentados com as amostras de todos os experimentos. O conjunto de dados para classificação consistiu em um total de 1680 amostras (120 amostras de cada um dos 14 experimentos) para cada um dos cinco microfones. Buscou-se com a classificação distinguir situações operacionais normais na tubulação de situações com vazamentos. Já os algoritmos de regressão foram usados para localizar os vazamentos. Para cada amostra fornecida, cabia aos modelos indicar a distância dos orifícios em relação a entrada da tubulação. Por esse motivo aos algoritmos de regressão só foram fornecidos os dados referentes aos experimentos com um único vazamento (experimentos 4 a 10 na Tabela 3). O conjunto de dados para regressão consistiu em 840 amostras (120 amostras para cada um dos sete experimentos) para cada um dos cinco microfones.

Do total de dados 80% deles foram separados para treino e 20% para validação dos modelos. O conjunto de dados que continha os componentes principais (dados nos quais foi aplicado o PCA) foi fornecido a todos os algoritmos de aprendizado de máquina. Em contrapartida, o conjunto de dados contendo os atributos mais importantes foi usado somente com três algoritmos: Floresta Aleatória, *Adaboost e Xgboost*; uma vez que esses três modelos têm como base árvores de decisão (Adaboost pode ser usado com outros modelos, não somente Árvores de Decisão), e por isso espera-se que compartilhem os atributos mais importantes. Um esquema simplificado do procedimento adotado é mostrado na Figura 18.

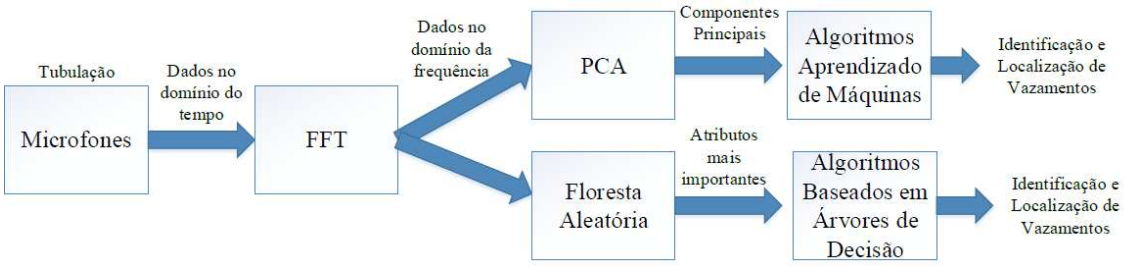

Figura 18 - Esquema do procedimento adotado para identificar e localizar vazamentos.

Cada um dos modelos empregados tem um certo número de hiperparâmetros que devem ser especificados pelo usuário, ou seja, eles não são aprendidos automaticamente dos dados durante a etapa de treinamento. Exemplos desses hiperparâmetros são o número de vizinhos mais próximos no modelo K vizinhos mais próximos, o número de neurônios e de camadas internas em uma rede neural etc. A seleção criteriosa desses hiperparâmetros é um passo essencial para que resultados satisfatórios sejam obtidos com os modelos. Para cada um deles um conjunto de hiperparâmetros foi selecionado (o restante foi deixado com seus valores padrão), e para cada um desses hiperparâmetros foi determinado um intervalo de valores a ser testado. Essas duas escolhas basearam-se no conhecimento prévio dos modelos e em considerações de custo computacional. Limitou-se assim a seleção aos hiperparâmetros com maior impacto no resultado e ao mesmo tempo restringiu-se o número de valores testados. Dessa forma reduziu-se o escopo da busca dentre as infinitas possibilidades que poderiam ser empregadas.

Após a seleção dos hiperparâmetros e dos seus valores, todas as combinações possíveis entre eles foram testadas. O treinamento foi conduzido empregando-se uma técnica conhecida como validação cruzada. Essa técnica consiste em se dividir os dados de treinamento em k grupos, cada um deles contendo a mesma quantidade de dados. O modelo é treinado com os dados contidos em k-1 grupos e validado com os dados do grupo restante. O processo é repetido até que os dados de cada grupo tenham sido usados uma vez para validação. Dessa forma, ao final do processo o modelo é treinado e validado com dados não usados no treinamento um total de k vezes. Isso garante confiabilidade nos resultados obtidos sem que seja necessária uma grande quantidade de dados. No presente trabalho o valor adotado para k foi cinco. Assim para cada combinação de valores dos hiperparâmetros os modelos foram treinados e validados cinco vezes. Um esquema simplificado da validação cruzada com k igual a cinco é mostrado na Figura 19.

Para comparação de performance dos modelos com as diferentes combinações de valores de hiperparâmetros foi usada a acurácia no caso dos algoritmos de classificação

e a raiz do erro quadrático médio (RMSE) para os algoritmos de regressão. A acurácia simplesmente mediu a efetividade média dos modelos, ou seja, a proporção dos elementos que foram corretamente classificados. Já o RMSE é a raiz do erro quadrático médio, onde para o caso aqui tratado o erro é a diferença entre o valor predito pelos modelos e o valor real das distâncias entre os orifícios e o início da tubulação. O RMSE é geralmente uma boa métrica, uma vez que é justamente essa métrica que a maioria dos modelos de regressão busca minimizar na etapa de treinamento. Com base nessas duas métricas foram selecionados os melhores valores para os hiperparâmetros. Esses valores foram adotados durante todo o trabalho (os valores para os hiperparâmetros de todos os modelos estão no Apêndice A). Os modelos foram então treinados com a totalidade dos dados de treinamento e posteriormente validados com os 20% dos dados que foram separados para validação.

Iteração 1	Teste	Treino	Treino	Treino	Treino
Iteração 2	Treino	Teste	Treino	Treino	Treino
Iteração 3	Treino	Treino	Teste	Treino	Treino
Iteração 4	Treino	Treino	Treino	Teste	Treino
Iteração 5	Treino	Treino	Treino	Treino	Teste

Figura 19 - Esquema da validação cruzada. Dados divididos em 5 grupos. Fonte: Adaptado de JAMES et al. (2013).

Duas outras métricas foram empregadas para avaliar os resultados dos modelos de classificação: precisão e sensibilidade. Para o cálculo dessas duas métricas os experimentos foram divididos em duas categorias: experimentos sem vazamentos (experimentos 1 a 3 na Tabela 3) e experimentos com vazamentos (experimentos 4 a 14 na Tabela 3). A precisão indica a taxa na qual os experimentos classificados como vazamentos provinham realmente de experimentos com vazamentos. Elevadas precisões correspondem a baixas taxas de falsos alarmes. Já a sensibilidade é a taxa com que todos os experimentos com vazamentos são classificados como vazamentos. Assim elevadas sensibilidades correspondem a elevadas taxas de identificação de vazamentos. A definição matemática dessas duas métricas é apresentada nas Equações 7 e 8. Nessas equações *tp* corresponde ao número de positivos verdadeiros, vazamentos sendo corretamente classificados como vazamentos; *fp* são falsos positivos, não vazamentos classificados como vazamentos; *fn* são falsos negativos, vazamentos classificados como não vazamentos.

$$Precisão = \frac{tp}{tp+fp} \quad (7)$$

$$Sensibilidade = \frac{tp}{tp + fn} \quad (8)$$

Já os modelos de regressão foram avaliados, além do RMSE, com a métrica erro de localização. Essa métrica é a fração entre o tamanho do erro cometido pelo modelo e a extensão da tubulação monitorada pelo sensor (Equação 9). Essa métrica foi adotada com a finalidade de se comparar os resultados obtidos neste trabalho com outros resultados disponíveis na literatura científica.

$$Erro\ de\ localização\ (\%) = \frac{valor\ predito - valor\ real}{extensão\ da\ tubulação\ sendo\ monitorada} \quad (9)$$

Por fim os modelos foram treinados com a totalidade de dados de alguns dos experimentos sendo deixados de fora do treinamento. Os algoritmos de classificação foram treinados com os três experimentos sem nenhum vazamento e com dados de seis dos sete experimentos com um único vazamento. Os experimentos com dois vazamentos simultâneos não foram empregados para reduzir o desbalanceamento entre o número de experimentos com e sem vazamentos. Foram conduzidos sete testes, de modo que cada um dos experimentos com um único vazamento foi deixado de fora da etapa de treinamento uma vez. Após cada treinamento os modelos classificaram os dados do experimento ausente no treinamento. A predição foi considerada correta quando a amostra foi classificada como um dos experimentos com vazamentos. Em seguida, o modelo com o melhor desempenho foi testado com mais de um experimento sendo deixado de fora do treinamento. O processo seguiu até que a etapa de treinamento continha dados de somente seis experimentos, três com vazamentos e três sem vazamentos. Processo semelhante foi realizado com os algoritmos de regressão, seis dos sete experimentos com um único vazamento foram usados no treinamento. O modelo foi então usado para predizer a distância para os dados do experimento que foi deixado de fora do treinamento. O valor predito foi comparado com a distância real do orifício em relação a entrada da tubulação.

5.Resultados e Discussões

Nessa seção serão expostos e discutidos os resultados obtidos no monitoramento da tubulação através de microfones. Os diversos experimentos, com e sem vazamentos, tiveram dados coletados no domínio do tempo. Esses dados foram convertidos para o domínio da frequência e em seguida sua dimensionalidade foi reduzida, para finalmente serem fornecidos a algoritmos de aprendizado de máquina, que por sua vez identificaram e localizaram vazamentos. Serão apresentados os resultados do monitoramento tanto no domínio do tempo quanto no domínio da frequência, e em seguida os resultados alcançados com os algoritmos de classificação e de regressão.

5.1 Monitoramento da Tubulação

Os gráficos da Figura 20 mostram o comportamento do sinal sonoro captado pelos microfones em dois experimentos: experimento 6 (orifício de 2 mm) em (a) e experimento 5 (orifício de 0,5mm) em (b). Os vazamentos foram provocados 10 s após o início de cada experimento. Pode-se ver em ambos os gráficos que o vazamento provoca uma mudança abrupta na amplitude do sinal sonoro. Em (a) nota-se com clareza que aproximadamente 10 s após o início do vazamento o sinal sonoro atinge um novo regime permanente, no qual as amplitudes absolutas médias (soma do valor absoluto das amplitudes) são superiores às do estado inicial no qual o vazamento não estava ocorrendo. Esse padrão se repetiu no sinal captado pelos cinco microfones, porém o aumento das amplitudes absolutas captadas diminuiu na medida em que o microfone estava mais afastado do orifício. A mudança nas amplitudes absolutas médias provocadas pelos vazamentos foi sutil em (b), a ponto de se tornar imperceptível, mesmo para o microfone mais próximo ao vazamento. Em (a) o valor da média das amplitudes absolutas antes do início do vazamento (intervalo de tempo entre t = 0 e t = 10 s) para o sinal captado pelo microfone 5 foi de 0,0105 V. Esse valor subiu para 0,204 V para o intervalo de tempo entre t= 20 s e t = 60 s, o que corresponde a um aumento de 1843%. Enquanto em (b) o mesmo valor variou de 0,00831 V para 0,0104 V, aumento de 25 %. Essa diferença nas duas variações ilustra o desafio que os pequenos vazamentos representam para os sistemas de detecção. Na medida em que orifícios de onde ocorrem a fuga de gás tornam-se menores, mais próximo se torna o comportamento do sinal acústico gerado pela tubulação em condições operacionais normais e com vazamentos.

Através da aplicação da Transformada Rápida de Fourier foi possível analisar os espectros do sinal captado pelos microfones nas diferentes condições dos experimentos.

Foram padrões presentes nesses espectros, identificados pelos algoritmos de aprendizado de máquina, que possibilitaram a identificação e localização dos vazamentos. A Figura 21 apresenta os espectros do sinal de diferentes experimentos capturados pelo microfone 5. Pode-se notar nos gráficos que todos os espectros apresentam diferenças entre si, os mais semelhantes são os 21 (a) e 21 (d), que são os espectros dos experimentos sem vazamento e com vazamento no orifício 2 (0,5 mm). Esses são espectros do sinal captado pelo microfone mais afastado dos orifícios. Consequentemente o sinal sofreu maior atenuação, o que tende a minimizar a diferença entre o sinal dos experimentos sem vazamentos e com vazamentos, especialmente para os pequenos orifícios. Para ilustrar essa atenuação os gráficos da Figura 22 mostram lado a lado os espectros do sinal captado pelos microfones 1 e 5 durante o experimento 4 (vazamento no orifício de 0,5 mm), onde pode-se notar as menores amplitudes do sinal captado pelo microfone 5 na comparação com o mesmo sinal captado pelo microfone 1.

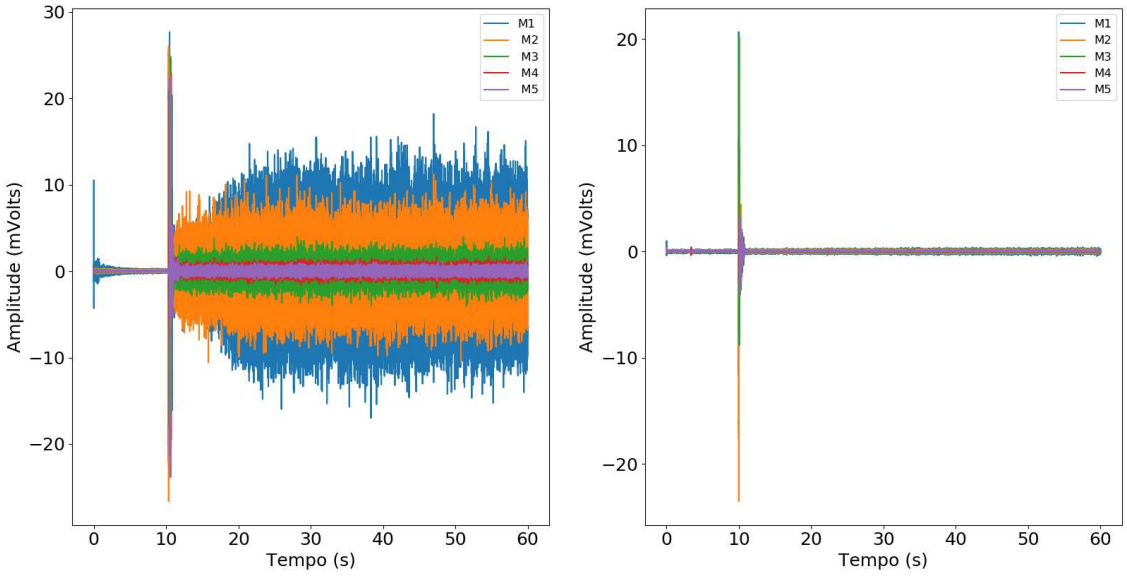

Figura 20 - Comportamento das amplitudes captadas pelos microfones durante os experimentos 5 (b) e 6 (a).

Já a Figura 23 mostra o espectro do sinal com dois vazamentos simultâneos, vazamentos nos orifícios com 1 mm e 2 mm (orifícios 1 e 3 respectivamente) e os espectros dos sinais dos dois vazamentos ocorrendo separadamente. Percebe-se que no espectro com os dois vazamentos predomina o vazamento de maior dimensão, assim o espectro da Figura 23 (c) se assemelha mais com o da Figura 23 (b) do que com o da Figura 23 (a).

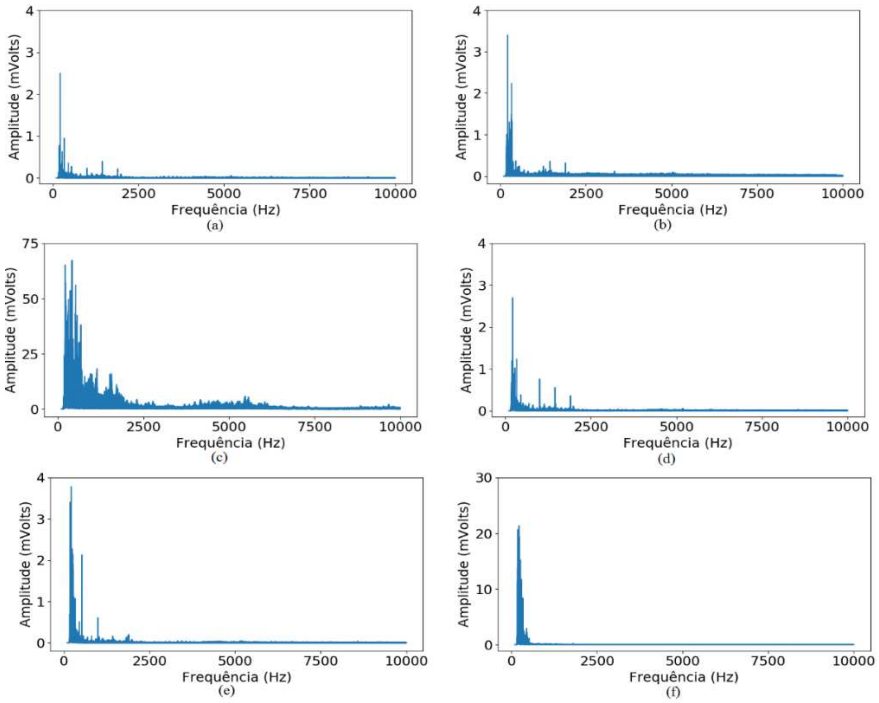

Figura 21 - Espectro do sinal captado pelo microfone 5 em diferentes experimentos. (a) Sem vazamentos. (b) Sem vazamentos com batidas no botijão. (c) Sem vazamentos com batidas na tubulação. (d) Vazamento no orifício 2 (0,5 mm). (e) Vazamento no orifício 1 (1 mm). (f) Vazamento no orifício 3 (2 mm).

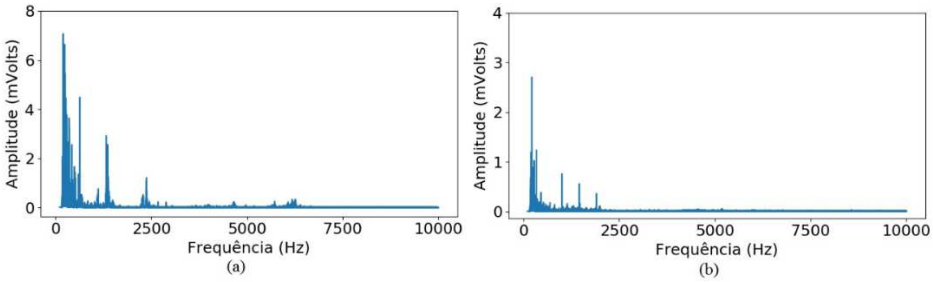

Figura 22 - Espectro do sinal captado durante o experimento 4 (orifício de 0.5 mm). (a) Microfone 1. (b) Microfone 5.

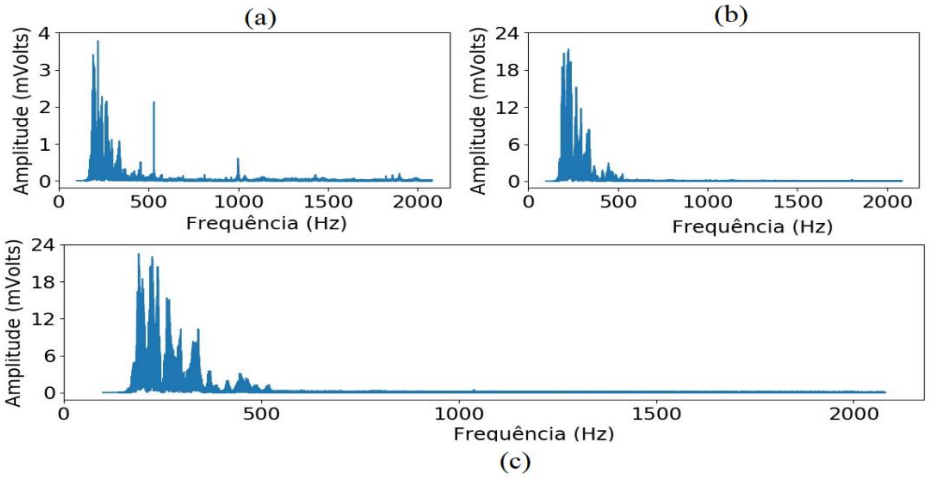

Figura 23 - Espectros de três experimentos. Sinal captado pelo microfone 5. (a) Experimento 4 (orifício de 1 mm). (b) Experimento 6 (orifício de 2mm). (c) Experimento 11. (Vazamentos nos orifícios de 1mm e de 2 mm simultaneamente).

5.2 Identificação de Vazamentos

Acerca da identificação de vazamento, nota-se que cinco dos oito modelos de aprendizado de máquina aplicados atingiram elevadas acurácias quando empregados para classificar os quatorze experimentos conduzidos (Tabela 4). As performances dos três algoritmos, aos quais foram fornecidos os dados contendo os atributos mais importantes selecionados pelo algoritmo Floresta Aleatória, foram claramente superiores. Por exemplo, o modelo *Adaboost* alcançou uma acurácia de 0,839 quando alimentado com os componentes principais. Já quando os atributos mais importantes foram empregados a acurácia chegou a 0,970, aumento de aproximadamente 15% (nos dois casos os resultados referem-se ao microfone 5).

Tabela 4 - Resultados obtidos com os algoritmos de classificação (dados de validação).

	Microfone 3			Microfone 5		
Modelo	Acurácia	Precisão	Sensibilidade	Acurácia	Precisão	Sensibilidade
KNN[a]	0,223	0,968	0,798	0,235	0,968	0,814
LR[a]	0,583	0,889	1,000	0,557	0,967	0,776
NN[a]	0,616	0,984	0,692	0,354	1,000	0,433
SVM - Linear[a]	0,753	0,996	1,000	0,863	0,974	0,992
SVM - RBF[a]	0,780	0,985	1,000	0,851	0,963	0,992
Floresta Aleatória	0,792	0,985	1,000	0,821	0,997	0,996
AdaBoost[a]	0,801	0,989	1,000	0,839	0,989	0,996
XgBoost[a]	0,789	0,992	0,996	0,842	0,992	0,996
Floresta Aleatória[b]	0,920	0,963	1,000	0,961	0,996	0,996
Adaboost[b]	0,964	0,981	1,000	0,970	0,996	1,000
Xgboost[b]	0,899	0,985	0,996	0,958	0,992	1,000

[a] – Dados fornecidos aos modelos continham os componentes principais dos dados originais

[b] – Dados fornecidos ao modelo continham os atributos mais importantes dos dados originais selecionado pelo algoritmo floresta aleatória.

Os melhores resultados foram obtidos com os dados captados pelo microfone 1, o mais próximo da entrada da tubulação, seguido do microfone 5, o mais próximo da extremidade final, enquanto a pior performance foi obtida com o microfone 3. Uma possível explicação para esse fato observado foi a localização dos microfones 1 e 5 nas extremidades da tubulação. Desse ponto em diante as análises serão feitas levando-se em conta os resultados obtidos com os dados coletados pelo microfone 5. A razão dessa escolha foi a localização desse microfone na posição mais relevante, a mais afastada dos orifícios. Pela mesma razão os resultados para o microfone 5 estão presentes na Tabela 4,

enquanto os do microfone 3 para apresentar os piores resultados obtidos. Os resultados para os demais microfones estão no Apêndice B (no Apêndice B estão todos os resultados obtidos com os algoritmos de classificação, inclusive os da validação cruzada).

Em relação ao problema de identificar vazamentos, a distinção que precisa ser feita é somente entre experimentos com vazamentos e experimentos sem vazamentos. Para esse fim as métricas precisão e sensibilidade foram empregadas. Considerou-se uma classificação binária, experimentos com vazamento e experimentos sem vazamentos. Assim, para o objetivo buscado no presente trabalho essas duas métricas têm mais a informar do que a acurácia. Os três experimentos sem vazamentos foram agrupados em uma única classe, e os onze experimentos com vazamentos na outra classe.

O modelo Floresta Aleatória, por exemplo, quando recebeu os dados contento os componentes principais atingiu uma precisão e uma sensibilidade de 0,997 e 0,996 respectivamente. A precisão indica que de todas as vezes que o modelo classificou uma amostra como um dos experimentos com vazamento, 99,7% dessas amostras realmente pertenciam experimentos com vazamentos. Ou seja, apenas 0,3% dos casos foram alarmes falsos, quando experimentos sem vazamentos foram classificados como vazamentos. Já a sensibilidade quer dizer que de todas as amostram provenientes de experimentos com vazamentos, 99,6% delas foram classificadas como um experimento com vazamento. Apenas 0,4% dos vazamentos não foram identificados pelo modelo. Para salientar esse resultado obtido, o de que poucos erros de classificação provieram de experimentos com vazamentos sendo classificados como experimentos sem vazamentos e vice-versa, na Figura 24 é apresentada a matriz de confusão da modelo Floresta Aleatória alimentado com os componentes principais. Nota-se na matriz que o experimento 4 (vazamento no orifício 1 (1 mm de diâmetro)) foi classificado erroneamente o maior número de vezes. Das 27 amostras desse experimento, treze foram corretamente classificadas, enquanto 5 amostras foram classificadas como experimento 7 (vazamento no orifício 4 (1 mm de diâmetro)) e 9 como experimento 11 (vazamentos nos orifícios 1 e 3 (1 mm e 2 mm de diâmetro respectivamente)). Ou seja, os erros provieram da distinção entre experimentos com vazamentos em orifícios do mesmo diâmetro. De todas as amostras de experimento sem vazamentos, apenas uma foi classificada como experimento com vazamento, experimento 2 (sem vazamentos com batidas no botijão) classificado como 7 (vazamento no orifício 4 (1 mm de diâmetro)). Uma única amostra de experimento com vazamento foi classificada incorretamente como sem vazamento,

experimento 8 (vazamento no orifício 5 (4 mm de diâmetro)) classificado como 2 (sem vazamentos com batidas no botijão).

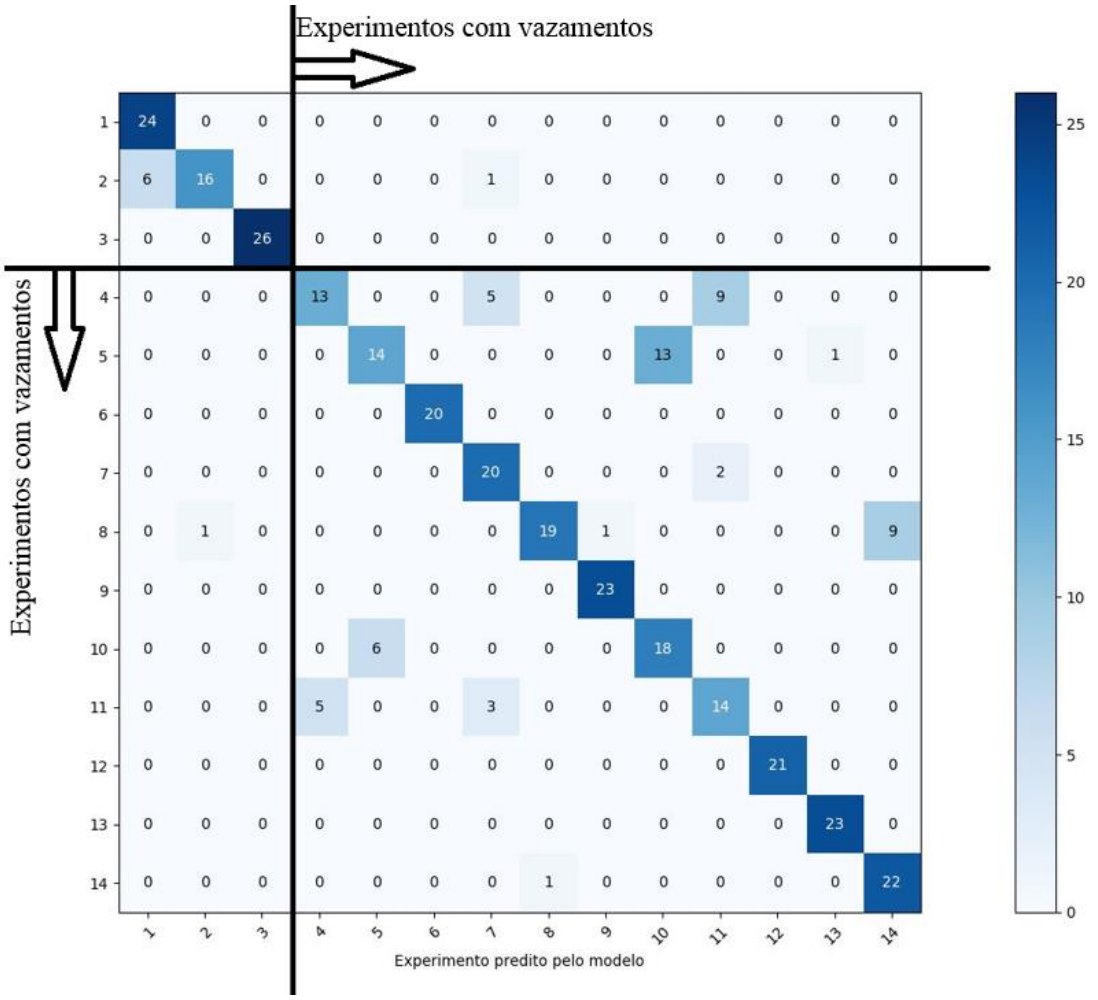

Figura 24 - Matriz de confusão do modelo Floresta Aleatória quando recebeu os dados contendo os componentes principais.

Os resultados apresentados até aqui foram alcançados quando foram apresentados aos modelos dados de todos os experimentos durante a etapa de treinamento. O passo seguinte foi testar a performance dos modelos com dados de experimentos que não estavam presentes no treinamento. Esse é um resultado crucial, uma vez que os modelos quando implementados no monitoramento de uma tubulação certamente lidarão com eventos desconhecidos. Contrariamente aos resultados anteriores, as melhores performances foram alcançadas quando os componentes principais foram empregados (Tabela 5). Dados de nove experimentos foram usados no treinamento, experimentos 1 a 3 e seis dos sete experimentos com um único vazamento. A predição foi considerada correta se o experimento foi classificado como um dos experimentos com vazamentos. Dentre os modelos o melhor resultado foi alcançado com o algoritmo Floresta Aleatória, que corretamente classificou quase a totalidade das amostras. A menor acurácia para esse

modelo (dados contendo os componentes principais) foi de 0,957, quando e experimento 7 foi deixado de fora do treinamento. Como os dados continham amostras de 3 experimentos sem vazamentos e 6 com vazamentos, um classificador aleatório teria chegado a um resultado de aproximadamente 0,667. Considerando-se que os dados classificados continham 120 amostras, 95,7% no mínimo corretamente classificadas é um resultado significativo mesmo na situação testada que continha um número desbalanceado de elementos das duas classes (a predição do modelo foi considerada correta quando a amostra foi classificada como um vazamento, mas os dados de treinamento continham 6 experimentos com vazamentos e 3 sem vazamentos).

Tabela 5 - Resultado da classificação de dados de um experimento que foi deixado de fora do treinamento.

Microfone 5							
	Acurácia quando o n-ésimo experimento foi deixado de for a do treinamento						
n / Modelo	3	4	5	6	7	8	9
KNN[a]	0,992	0,058	0,858	0,983	0,992	1,000	0,067
LR[a]	0,992	0,000	0,983	1,000	0,933	1,000	0,125
NN[a]	0,808	0,183	1,000	1,000	0,775	0,000	0,900
SVM - Linear[a]	1,000	0,817	1,000	1,000	1,000	1,000	0,817
SVM - RBF[a]	1,000	0,800	1,000	1,000	0,008	1,000	0,883
Floresta aleatória[a]	1,000	0,992	0,967	1,000	0,957	1,000	0,992
AdaBoost[a]	0,967	0,967	0,950	1,000	0,992	1,000	0,983
XgBoost[a]	0,992	0,975	0,742	0,992	0,983	0,958	0,983
Floresta aleatória[a]	1,000	0,475	0,158	1,000	0,000	0,692	1,000
Adaboost[b]	1,000	0,458	0,023	0,983	0,175	0,750	1,000
Xgboost[b]	1,000	0,458	0,550	0,917	0,208	0,850	1,000

[a] — Dados fornecidos aos modelos continham os componentes principais dos dados originais

[b] — Dados fornecidos ao modelo continham os atributos mais importantes dos dados originais selecionado pelo algoritmo floresta aleatória.

Como os melhores resultados na classificação de dados desconhecidos foram alcançados com o algoritmo Floresta Aleatória, esse modelo foi testado como mais de um experimento sendo deixado de fora do treinamento. Foram empregados os dados contendo os componentes principais. O processo começou com dados de dois experimentos sendo deixados de fora, e prosseguiu até que no treinamento restaram dados de apenas seis experimentos. Os resultados variaram de um mínimo de 0,506 até todas as amostram sendo corretamente classificadas, dependendo dos experimentos que foram apresentados ao modelo no treinamento (Tabela 6). Esse resultado mostra que os modelos

devem ser treinados com dados de experimentos que sejam os mais variados, ou seja, vazamentos que cubram uma grande amplitude de diâmetros, de modo que quando um evento desconhecido acontecer, o comportamento sonoro do sistema terá semelhança com alguma das situações que foram empregadas no treinamento, permitindo assim a correta classificação.

Tabela 6 - Resultados obtidos com o modelo Floresta Aleatória quando dados de mais de um experimento foi deixado de foram do treinamento.

Experimentos não presentes no treinamento	Microfone 3	Microfone 5
	Acurácia	
6,8	0,971	1,000
5,6	0,825	1,000
8,9	0,983	1,000
6,8,9	0,981	1,000
3,4,5	0,867	1,000
4,6,8	0,986	1,000
6,7,8,9	0,765	0,967
3,6,8,9	0,504	0,506
3,4,7,8	0,758	0,973
3,4,5,7	0,888	1,000

Considerando-se todos os resultados apresentados até aqui, os componentes principais dos dados no domínio da frequência devem ser usados. O uso dos dados contendo os atributos mais importantes levou a acurácias mais elevadas quando dados de todos os experimentos estavam presentes no treinamento, porém isso não se repetiu nos testes seguintes, quando dados foram deixados de fora do treinamento. Assim o critério para a escolha foi a generalização, a classificação de experimentos desconhecidos. Esse critério indica a maior aplicabilidade ao monitoramento de tubulações em situações não experimentais, quando os modelos certamente serão expostos a situações desconhecidas. O mesmo critério recomenda a adoção do modelo Floresta Aleatória. A performance desse modelo foi pouco inferior ao do modelo *Adaboost*, o de melhor performance, nos primeiros testes (Tabela 4). Porém na etapa mais importante (classificação de dados desconhecidos) a performance do algoritmo Floresta Aleatória foi claramente superior aos demais (Tabela 5).

A importância dos resultados obtidos reside na capacidade demonstrada pelo sistema proposto em detectar pequenos vazamentos em tubulações operando em baixas

pressões, uma vez que esse é um problema ainda não solucionado. SANTOS et al. (2014), por exemplo, propuseram uma técnica que faz uso de microfones e redes neurais artificiais para identificar vazamentos e predizer a magnitude deles. No experimento o método não foi capaz de detectar o menor dos vazamentos (1 mm de diâmetro) quando a pressão na tubulação era de 392 kPa. Quando a pressão foi aumentada para 588 kPa, todos os vazamentos foram identificados. MENG et al. (2012) aplicaram sensores de pressão dinâmica e um conjunto de características do sinal captado para identificar vazamentos. Os autores reportaram baixas performances da técnica proposta quando empregada para detectar fuga de gás de um orifício de 0,45 mm quando a tubulação estava submetida a uma pressão de 1600 kPa. Segundo os autores resultados melhores foram obtidos na medida que pressões superiores foram aplicadas. LIU et al. (2017) empregou sensores de pressão dinâmica em conjunto com leis de propagação de ondas. A técnica proposta pelos autores teve sucesso na detecção de vazamentos na tubulação operando em pressões superiores a 390 kPa. Com base nesses exemplos, e em outros mais disponíveis na literatura científica, comprova-se que a detecção de pequenos vazamentos em tubulações que operam em baixas pressões consiste em um enorme desafio. Esse desafio vem do fato de que na medida em a cai a pressão, o sinal acústico gerado pela tubulação operando com vazamentos torna-se mais semelhante ao sinal gerado em condições operacionais normais. A semelhança entre os dois sinais não somente dificulta a detecção dos vazamentos, mas também aumenta a ocorrência de falsos alarmes, nos quais durante a operação normal o sistema de detecção acusa a ocorrência de um vazamento.

Dessa forma a técnica proposta no presente trabalho representa uma atrativa alternativa para o monitoramento de tubulações que operam em baixas pressões. Durante os experimentos o método proposto foi capaz de identificar todos os vazamentos testados, mesmo quando o gás escapou de orifícios de apenas 0,5 mm de diâmetro. Durante todos os testes a pressão na tubulação era de 100 kPa. Tão importante quanto a detecção dos vazamentos, foi a baixa taxa de falsos alarmes. Essa baixa taxa foi alcançada mesmo na presença de perturbações externas (batidas na tubulação e batidas no vaso), fatores comuns na operação de tubulações e que acabam por confundir sistemas de detecção de vazamentos.

5.3 Localização de Vazamentos

Na localização de vazamentos, algoritmos de regressão foram empregados para calcular a localização dos orifícios na tubulação. Nessa etapa foram empregados somente os dados de experimentos com um único vazamento (experimentos 4 a 10 na Tabela 3). Para a comparação dos resultados a métrica empregada foi a raiz do erro quadrático médio (RMSE). Igualmente ao que ocorreu com os algoritmos de classificação, os melhores resultados foram obtidos com os modelos de regressão alimentados com os dados contendo os atributos mais importantes selecionados pelo algoritmo Floresta Aleatória. Com o algoritmo *Adaboost*, quando alimentado pelos componentes principais, a RMSE alcançada foi de 1,32 m, já quando foram empregados os atributos mais importantes com o mesmo algoritmo a RMSE foi de 0,14 m, queda de aproximadamente 89%. No comparativo entre os modelos, a menor RMSE foi alcançada com o uso do algoritmo *Adaboost* (Tabela 7). O pequeno valor da RMSE significa que os valores preditos pelo modelo são muito próximos as reais posições dos orifícios (Tabela 8). O maior erro de localização foi de 0.42%, enquanto a média dos erros de localização para as sete posições foi de apenas 0.1%. Os resultados para os demais microfones estão no Apêndice C (no Apêndice C estão os resultados completos obtidos com os algoritmos de regressão, inclusive na validação cruzada).

Tabela 7 - Resultados obtidos com os algoritmos de regressão.

	Microfone 3	Microfone 5
Modelo	RMSE (m)	
KNN[a]	2,11	1,85
NN[a]	1,40	1,41
SVM - Linear[a]	1,77	1,78
SVM - RBF[a]	1,75	1,67
Florestas Aleatórias[a]	1,42	1,21
AdaBoost[a]	2,02	1,32
XgBoost[a]	1,54	1,21
Florestas Aleatórias[b]	0,96	0,35
AdaBoost[b]	0,54	0,14
XgBoost[b]	0,89	0,36

[a] – Dados fornecidos aos modelos continha os componentes principais dos dados originais

[b] – Dados fornecidos ao modelo continham os atributos mais importantes dos dados originais selecionado pelo algoritmo floresta aleatória.

Os algoritmos de regressão também foram testados com dados que não estavam presentes durante o treinamento. Dados de seis dos sete experimentos com um vazamento

foram usados no treinamento. Os modelos foram treinados no total cinco vezes. Em cada um desses treinamentos dados de um experimento diferente foram deixados de fora. Dados de dois experimentos nunca foram retirados, os dados dos experimentos nos quais ocorreram vazamentos nos orifícios mais próximo e mais afastado da entrada da tubulação. A não retirada dos dados desses dois experimentos se deu pela incapacidade dos algoritmos de aprendizado de máquina aplicados para extrapolações, ou seja, os modelos não são capazes de lidar com situações que estejam além daqueles presentes no treinamento. Assim, para o caso tratado nesse trabalho não se pode esperar predições adequadas para a posição dos orifícios cujos valores reais sejam ou maiores ou menores que todos os exemplos apresentados aos modelos no treinamento.

Tabela 8 - Valores reais e os preditos pelo modelo Adaboost para as posições dos orifícios na tubulação. Os dados usados continham os atributos mais importantes.

Posição Real (m)	Posição Predita (m)	Erro de Localização (%)
0,2	0,27	0,15
1	1,06	0,13
1,6	1,60	0,00
2,4	2,40	0,00
3,4	3,40	0,00
4,8	4,80	0,00
5,6	5,41	0,42

Conforme era de se esperar, os erros de localização para os dados não presentes no treinamento foram superiores aos obtidos anteriormente (Tabelas 7 e 8). Dentre todos os modelos testados, o *Xgboost* com os dados contendo os atributos mais importantes obteve o melhor desempenho (Tabela 9). O melhor resultado corresponde a um erro de localização médio de 1,75%. Além disso, para três das cinco predições os erros de localização foram iguais ou inferiores a 1%, enquanto os maiores e menores erros foram de 4,31% e 0,02%. Na comparação com os resultados obtidos quando dados de todos os experimentos foram apresentados no treinamento o erro de localização médio subiu de 0,1% para 1,75%, enquanto o maior erro de localização subiu de 0,42% para 4,31% (a comparação foi entre os melhores resultados obtidos). A causa desses erros foi a variação do sinal sonoro não somente com a distância em relação aos microfones, mas também com o diâmetro do orifício. Por exemplo, dado um sinal acústico captado com uma elevada amplitude para uma certa frequência. Esse sinal poderia ter sido provocado por

um vazamento em um orifício grande e afastado do microfone ou de um orifício menor e mais próximo do microfone.

Tabela 9 - - Predições para dados de experimentos que não foram usados no treinamento.

	Valor Real= 1m		Valor Real= 1.6 m		Valor Real= 2.4 m		Valor Real= 3.4 m		Valor Real= 4.8 m	
Modelo	Valor Predito (m)	Erro de Localização (%)	Valor Predito (m)	Erro de Localização (%)	Valor Predito (m)	Erro de Localização (%)	Valor Predito (m)	Erro de Localização (%)	Valor Predito (m)	Erro de Localização (%)
KNN[a]	5,6	9,68	1,26	0,72	2,85	0,95	3,4	0,00	1,6	6,74
NN[a]	5,6	9,68	3,26	3,49	0,44	4,13	1,06	4,93	2,15	5,58
SVM – Linear[a]	2,16	2,44	2,65	2,21	2,06	0,72	6,97	7,52	2,03	5,83
SVM – RBF[a]	4,69	7,77	0,43	2,47	0,63	3,72	4,75	2,85	3,67	2,38
Floresta Aleatória[a]	5,57	9,61	2,25	1,38	0,29	4,43	4,79	2,92	1,90	6,10
AdaBoost[a]	5,60	9,68	3,39	3,77	0,29	4,44	4,80	2,95	1,79	6,33
XgBoost[a]	5,57	9,61	2,30	1,48	0,24	4,55	4,71	2,75	1,71	6,50
Floresta Aleatória[b]	4,93	8,27	4,00	5,05	1,05	2,84	4,77	2,88	2,90	4,00
Adaboost[b]	5,6	9,68	4,8	6,74	0,71	3,56	4,8	2,95	1,60	6,74
Xgboost[b]	0,91	0,20	1,61	0,02	1,90	1,06	1,91	3,14	2,75	4,31

[a] — Dados fornecidos aos modelos continham os componentes principais dos dados originais

[b] — Dados fornecidos ao modelo continham os atributos mais importantes dos dados originais selecionado pelo algoritmo floresta aleatória

Muitas técnicas já foram propostas com a finalidade de localizar vazamentos em tubulações. Alguns resultados recentemente publicados na literatura científica são apresentados na Tabela 10. Todas as metodologias presentes nesses artigos requerem a criação de modelos que descrevam a velocidade de propagação do gás no interior da tubulação (técnicas baseadas em diferença de tempo para um o sinal gerado por vazamentos atingir dois sensores e técnicas baseadas em diferença de velocidade) ou em modelos que descrevam a propagação das ondas sonoras na tubulação (técnicas baseadas na atenuação das ondas sonoras). É possível atingir elevadas precisões com esses modelos, porém eles são altamente específicos, complicados e ainda requerem a medição de outras variáveis, como a temperatura, densidade, pressão etc. Além disso, a maioria das técnicas propostas foram aplicadas em tubulações operando em altas pressões. Quanto mais alta a pressão, mais distinto o sinal gerado por um vazamento comparado as condições operacionais normais, facilitando assim as medidas tanto das amplitudes do sinal quando das diferenças de tempo ou velocidade, o que por fim acaba por reduzir os erros de localização.

A técnica proposta por LI et al. (2016) foi aplicada em uma tubulação submetida a baixas pressões. Os autores reportaram um erro de localização máximo de 3,06% e um erro de localização médio de 1,49%. Esses resultados são próximos aos obtidos no presente trabalho, que foram de 4,32% e 1,75% respectivamente.

A metodologia proposta no presente trabalho foi capaz de identificar a localização dos vazamentos. Trata-se de uma alternativa atrativa para o monitoramento de tubulações que operam a baixas pressões, pois é mais simples de ser implementada no comparativo com outras técnicas, além de apresentar a vantagem de combinar as tarefas de identificação e localização dos vazamentos.

Tabela 10 - Metodologias propostas para a localização de vazamentos.

Erro de localização máximo (%)	Técnica	Pressão (kPa)	Referência
3,06	Sensores de emissão acústica, diferença de tempo, correlação cruzada aprimorada	100	(LI et al., 2016)
19,38	Sensores de emissão acústica, Correlação cruzada	100	(LI et al., 2016)
10,7	Sensores de emissão acústica, características do sinal acústico	300	(CUI et al., 2016)
0,725	Sensores de pressão dinâmica, transformada wavelet, modelo de propagação de ondas	1000	(LIU et al., 2018)
0,735	Sensores de pressão dinâmica, transformada wavelet, diferença de tempo	1000	(LIU et al., 2018)
0,874	Sensores de pressão dinâmica, Diferença de tempo	1200	(LIU et al., 2017b)
0,59	Sensores de pressão dinâmica, Diferença de tempo	3000	(LIU et al., 2019b)
2,44	Sensores de pressão dinâmica, diferença de velocidade	3000	(LIU et al., 2019b)
1,83	Sensores de pressão dinâmica, atenuação de amplitude	3000	(LIU et al., 2019b)
11,68	Sensores de pressão dinâmica, atenuação de amplitude	3000	(LIU et al., 2019b)
1,48	Sensores de pressão dinâmica, diferença de tempo	5000	(LIU et al., 2019a)
0,59	Sensores de pressão dinâmica, diferença de velocidade	5000	(LIU et al., 2019a)

6.Conclusões e perspectivas para trabalhos futuros

A técnica que combina o método acústico (microfones) com algoritmos de aprendizado de máquina proposta no presente trabalho foi eficiente na detecção e localização de vazamentos.

Os experimentos mostraram que os vazamentos provocam uma abrupta elevação da amplitude do sinal sonoro captado pelos microfones, e alguns segundo após o início dos vazamentos as amplitudes se estabilizam em um novo patamar, cujo valores absolutos são superiores aos do estado inicial. Os valores atingidos pelas amplitudes dependem tanto do diâmetro do vazamento, a relação é direta, quanto da distância entre o orifício e o microfone, nesse caso a relação é inversa.

O sinal capturado pelos microfones foi convertido do domínio do tempo para o domínio da frequência. Os espectros dos experimentos com vazamentos mostraram-se diferentes dos experimentos sem vazamento. A maior semelhança for entre o espectro do experimento com vazamento no orifício de 0,5 mm e o experimento sem vazamentos, motivo pelo qual a detecção de pequenos vazamentos é desafiadora.

Com a aplicação da Análise de Componentes Principais (PCA) e a Extração dos Atributos mais Importantes do algoritmo Floresta Aleatória foi possível reduzir substancialmente o número de variáveis empregadas no treinamento dos algoritmos de aprendizado de máquina. Com o uso do PCA as 3340 amplitudes no domínio da frequência foram reduzidas a até 81 componentes principais, enquanto a Extração dos Atributos mais importantes selecionou um mínimo 154 atributos.

Os melhores resultados na detecção dos vazamentos foram alcançados com o uso do modelo Floresta Aleatória alimentado com os componentes principais do conjunto de dados. Esse algoritmo foi capaz de identificar corretamente 99,6% dos experimentos com vazamentos, mesmo com o gás escapando em alguns dos casos de orifícios de apenas 0,5 mm de diâmetro. Tão importante quanto a capacidade de detectar os vazamentos, foi a baixa taxa de falsos alarmes alcançada, apenas 0,3 % das amostras de experimentos sem vazamentos foram classificadas como de experimentos com vazamentos. Essa baixa taxa de falsos alarmes foi alcançada na presença de perturbações externas. Dentre os experimentos executados, dois deles ocorreram sem vazamentos, mas com perturbações externas, batidas na tubulação e batidas no botijão de gás. Esse tipo de perturbação é comum na operação de tubulações, e tendem a ser confundidas com vazamentos, ou seja, aumentam a taxa de falsos alarmes. Quanto a classificação de experimentos

desconhecidos, não apresentados ao modelo no treinamento, o modelo atingiu uma taxa mínima de acerto de 95,7 %. Esse último resultado é essencial e mostra o sucesso da técnica proposta, uma vez que no monitoramento de tubulações o sistema lidará certamente com situações desconhecidas.

Já com relação a localização dos vazamentos a técnica alcançou boas estimativas. As melhores predições foram obtidas com o modelo Xgboost alimentado com os atributos mais importantes. O maior erro de localização com esse modelo foi de 4,32%, enquanto o erro de localização médio para cinco posições foi de 1,75%. Essas estimativas foram para dados totalmente desconhecidos do modelo.

Os resultados obtidos qualificam a técnica proposta como uma alternativa viável para monitorar tubulações de baixas pressões. É um método que emprega sensores de baixo custo e que combina tanto a detecção quanto a localização de vazamentos. Sendo assim ela pode ser empregada no monitoramento de tubulações como as redes de distribuição de gás ao consumidor doméstico e muitas outras presentes tanto em ambientes comerciais quanto industriais.

6.1. Sugestões para trabalhos futuros

Para dar continuidade a este trabalho sugere-se:

- Implementar e verificar o comportamento da técnica no monitoramento da tubulação em tempo real;
- Perfurar mais orifícios na tubulação, de forma que estejam mais bem distribuídos ao longo da estrutura;
- Verificar o desempenho da técnica com orifícios ainda menores;
- Verificar o desempenho da técnica sob pressões mais baixas;
- Envolver a tubulação com algum tipo de material a fim de reproduzir uma situação operacional real da tubulação. Esse tipo de cobertura pode gerar maior atenuação do sinal acústico.

Referências

ABDULSHAHEED, A.; MUSTAPHA, F.; GHAVAMIAN, A. A pressure-based method for monitoring leaks in a pipe distribution system: A Review. **Renewable and Sustainable Energy Reviews**, v. 69, n. January 2016, p. 902–911, 2017.

ADEC. Technical review of leak detection technologies. **Environmental Conservation**, v. I, p. 1–31, 1999.

ADNAN, N. F.; GHAZALI, M. F.; AMIN, M. M.; HAMAT, A. M. A. Leak detection in gas pipeline by acoustic and signal processing - A review. **IOP Conference Series: Materials Science and Engineering**, v. 100, p. 012013, 2015.

ALJAROUDI, A.; KHAN, F.; AKINTURK, A.; HADDARA, M.; THODI, P. Risk assessment of offshore crude oil pipeline failure. **Journal of Loss Prevention in the Process Industries**, v. 37, p. 101–109, 1 set. 2015.

CHO, J.; KIM, H.; GEBRESELASSIE, A. L.; SHIN, D. Deep neural network and random forest classifier for source tracking of chemical leaks using fence monitoring data. **Journal of Loss Prevention in the Process Industries**, v. 56, n. February, p. 548–558, 2018.

CUI-WEI, L.; YU-XING, L.; JUN-TAO, F.; GUANG-XIAO, L. Experimental study on acoustic propagation-characteristics-based leak location method for natural gas pipelines. **Process Safety and Environmental Protection**, v. 96, p. 43–60, 2015.

CUI, X.; YAN, Y.; MA, Y.; MA, L.; HAN, X. Localization of CO2 leakage from transportation pipelines through low frequency acoustic emission detection. **Sensors and Actuators, A: Physical**, v. 237, p. 107–118, 2016.

DI BLASI, M.; MURAVCHIK, C. Leak Detection in a Pipeline Using Modified Line Volume Balance and Sequential Probability Tests. **Journal of Pressure Vessel Technology**, v. 131, n. 2, p. 021701, 2009.

DOORHY, J. Real-Time Pipeline Leak Detection And Location Using Volume Balancing. **Pipeline & Gas Journal**, v. 238(2), p. 65–66, 2011.

DUDIĆ, S.; IGNJATOVIĆ, I.; ŠEŠLIJA, D.; BLAGOJEVIĆ, V.; STOJILJKOVIĆ, M. Leakage quantification of compressed air using ultrasound and infrared thermography.

Measurement, v. 45, n. 7, p. 1689–1694, 2012.

EL-ZAHAB, S.; MOHAMMED ABDELKADER, E.; ZAYED, T. An accelerometer-based leak detection system. **Mechanical Systems and Signal Processing**, v. 108, p. 58–72, 2018.

ERSHOV, O.; KLIMOV, A.; VAVILOV, V. Airborne Detection of Gas Leaks from Transmission Pipelines by Using a Laser System Operating in Visual , Near-IR , and Mid-IR Wavelength Bands. **System**, n. Figure 1, 2007.

FERNANDES, L. B.; SANTOS, R. B.; FILETI, A. M. F. **Principal Component Analysis in Multivariate Microphones Response to Simulated Leakage in Metal Pipeline of Compressed Air**. Modelling, Simulation and Identification / 841: Intelligent Systems and Control. **Anais**...Calgary,AB,Canada: ACTAPRESS, 31 ago. 2016Disponível em: <http://www.actapress.com/PaperInfo.aspx?paperId=456244>. Acesso em: 16 maio. 2019

GEIGER, G. State-of-the-art in leak detection and localization. **Pipeline Technology 2006 Conference**, v. 32, n. 4, p. 193, 2006.

GÉRON, A. **Hands On.Machine.Learning.with.Scikit-Learn.and.TensorFlow.**1. ed. O'Reilly Media,2017.

GUENTHER, T.; KROLL, A. Automated detection of compressed air leaks using a scanning ultrasonic sensor system. **SAS 2016 - Sensors Applications Symposium, Proceedings**, p. 116–121, 2016.

HASTIE, T.; TIBSHIRANI, R.; FRIEDMAN, J. The Elements of Statistical Learning The Elements of Statistical Learning. 2017.

HAYKIN, S. **Neural Networks and Learning Machines**. 3. ed. Pearson,2018.

HAYKIN, S.; ENGEL, P. M. **Redes neurais : princípios e prática**.2. ed.Bookman, 2001.

HOU, Q.; REN, L.; JIAO, W.; ZOU, P.; SONG, G. An Improved Negative Pressure Wave Method for Natural Gas Pipeline Leak Location Using FBG Based Strain Sensor and Wavelet Transform. v. 2013, 2013.

INAUDI, D.; BONT, R. Fast Detection and Localization of Small Ammonia Leaks Using

Distributed Fiber Optic Sensors. p. 91–100, 2013.

JACKSON, R. B.; DOWN, A.; PHILLIPS, N. G.; ACKLEY, R. C.; COOK, C. W.; PLATA, D. L.; ZHAO, K. Natural Gas Pipeline Leaks Across Washington, DC. **Environmental Science & Technology**, v. 48, n. 3, p. 2051–2058, 2014.

JAIN, A. K.; MURTY, M. N.; FLYNN, P. J. Data clustering: a review. **ACM Computing Surveys**, v. 31, n. 3, p. 264–323, 1 set. 1999.

JAMES, G.; WITTEN, D.; HASTIE, T.; TIBSHIRANI, R. **An Introduction to Statistical Learning**. [s.l: s.n.]. v. 103

JIN, H.; ZHANG, L.; LIANG, W.; DING, Q. Integrated leakage detection and localization model for gas pipelines based on the acoustic wave method. **Journal of Loss Prevention in the Process Industries**, v. 27, n. 1, p. 74–88, 2014.

JOBS, Q.; FOLLOW, R.; SOCIETY, T. I.; ENGINEERING, O.; DIAZ, A.; THOMAS, B.; CASTILLO, P.; GROSS, B.; HIDE, F. M. Active stand-off detection of gas leaks using an open- path quantum cascade laser sensor in a backscatter configuration. v. 05013, p. 1–2, 2016.

JOLLIFFE, I. T. Principal Component Analysis. Second Edition. **Springer Series in Statistics**, v. 98, p. 487, 2002.

KAYAALP, F.; ZENGIN, A.; KARA, R.; ZAVRAK, S. Leakage detection and localization on water transportation pipelines: a multi-label classification approach. **Neural Computing and Applications**, v. 28, n. 10, p. 2905–2914, 2017.

KENNEDY, J. L. **Oil and Gas Pipeline Fundamentals**. 2. ed. Tulsa, USA: [s.n.].

KHALIFA, A. E.; CHATZIGEORGIOU, D. M.; YOUCEF-TOUMI, K.; KHULIEF, Y. A.; BEN-MANSOUR, R. **Quantifying acoustic and pressure sensing for in-pipe leak detection**. ASME International Mechanical Engineering Congress and Exposition, Proceedings (IMECE). **Anais**...2010

LI, S.; ZHANG, J.; YAN, D.; WANG, P.; HUANG, Q.; ZHAO, X.; CHENG, Y.; ZHOU, Q.; XIANG, N.; DONG, T. Leak detection and location in gas pipelines by extraction of cross spectrum of single non-dispersive guided wave modes. **Journal of Loss Prevention in the Process Industries**, v. 44, p. 255–262, 2016.

LIU, C.; CUI, Z.; FANG, L.; LI, Y.; XU, M. Leak localization approaches for gas pipelines using time and velocity differences of acoustic waves. **Engineering Failure Analysis**, v. 103, n. April 2019, p. 1–8, 2019a.

LIU, C.; LI, Y.; FANG, L.; HAN, J.; XU, M. Leakage monitoring research and design for natural gas pipelines based on dynamic pressure waves. **Journal of Process Control**, v. 50, p. 66–76, 2017a.

LIU, C.; LI, Y.; FANG, L.; XU, M. Experimental study on a de-noising system for gas and oil pipelines based on an acoustic leak detection and location method. **International Journal of Pressure Vessels and Piping**, v. 151, p. 20–34, 2017b.

LIU, C.; LI, Y.; FANG, L.; XU, M. New leak-localization approaches for gas pipelines using acoustic waves. **Measurement: Journal of the International Measurement Confederation**, v. 134, p. 54–65, 2019b.

LIU, C.; WANG, Y.; LI, Y.; XU, M. Experimental study on new leak location methods for natural gas pipelines based on dynamic pressure waves. **Journal of Natural Gas Science and Engineering**, v. 54, n. October 2017, p. 83–91, 2018.

LU, Z.; SHE, Y.; LOEWEN, M. A Sensitivity Analysis of a Computer Model-Based Leak Detection System for Oil Pipelines. **Energies**, v. 10, n. 8, p. 1226, 2017.

LYONS, R. G. **Understanding digital signal processing**. [s.l.] Prentice Hall, 2011.

MENG, L.; YUXING, L.; WUCHANG, W.; JUNTAO, F. Experimental study on leak detection and location for gas pipeline based on acoustic method. **Journal of Loss Prevention in the Process Industries**, v. 25, n. 1, p. 90–102, 2012.

MOSTAFAPOUR, A.; DAVOODI, S.; GHAREAGHAJI, M. Acoustic emission source location in plates using wavelet analysis and cross time frequency spectrum. **Ultrasonics**, v. 54, n. 8, p. 2055–2062, 2014.

MURVAY, P. S.; SILEA, I. A survey on gas leak detection and localization techniques. **Journal of Loss Prevention in the Process Industries**, v. 25, n. 6, p. 966–973, 2012.

NUNES, I.; SPATTI, H.; FLAUZINO, R. A. **Redes Neurais Artificiais Para Engenharia e Ciências Aplicadas**. 2° ed. São Paulo: [s.n.].

OLIPHANT, T. E. Python for Scientific Computing. **Computing in Science & Engineering**, v. 9, n. 3, p. 10–20, 1 maio 2007.

PEDREGOSA, F.; VAROQUAUX, G.; GRAMFORT, A.; MICHEL, V.; THIRION, B.; GRISEL, O.; BLONDEL, M.; PRETTENHOFER, P.; WEISS, R.; DUBOURG, V.; VANDERPLAS, J.; PASSOS, A.; COURNAPEAU, D.; BRUCHER, M.; PERROT, M.; DUCHESNAY, É. Scikit-learn: Machine Learning in Python. **Journal of Machine Learning Research**, v. 12, n. Oct, p. 2825–2830, 2011.

PHILLIPS, N. G.; ACKLEY, R.; CROSSON, E. R.; DOWN, A.; HUTYRA, L. R.; BRONDFIELD, M.; KARR, J. D.; ZHAO, K.; JACKSON, R. B. Mapping urban pipeline leaks: Methane leaks across Boston. **Environmental Pollution**, v. 173, p. 1–4, 2013.

PICHLER, K.; LUGHOFER, E.; PICHLER, M.; BUCHEGGER, T.; KLEMENT, E. P.; HUSCHENBETT, M. Fault detection in reciprocating compressor valves under varying load conditions. **Mechanical Systems and Signal Processing**, v. 70–71, p. 104–119, 2016.

RUIZ, M.; MUJICA, L. E.; MUJÍA, J. M. COMPUTATIONAL STATISTICAL MONITORING OF OF HYDROCARBON TRANSPORTATION LINES. **7th ECCOMAS Thematic Conference on Smart Structures and Materials**, p. 1–14, 2015.

SANTOS, A.; YOUNIS, M. A sensor network for non-intrusive and efficient leak detection in long pipelines. **IFIP Wireless Days**, v. 1, n. 1, 2011.

SANTOS, R. B.; DE SOUSA, E. O.; DA SILVA, F. V.; DA CRUZ, S. L.; FILETI, A. M. F. Detection and on-line prediction of leak magnitude in a gas pipeline using an acoustic method and neural network data processing. **Brazilian Journal of Chemical Engineering**, v. 31, n. 1, p. 145–153, 2014.

SCOTT, S; BARRUFET, M. (TEXAS A & M. U. Worldwide Assessment of Industry Leak Detection Capabilities for Single & Multiphase Pipelines Project Report Prepared for the Minerals Management Service Under the MMS / OTRC Cooperative Research Agreement Task Order 18133. p. 125, 2003.

SLR. Assessing Risk and Modeling a Sudden Gas Release Due to Gas Pipeline Ruptures. p. 5–22, 2009.

SMITH, S. W. **The scientist and engineer's guide to digital signal processing**. [s.l.] California Technical Pub, 1997.

SOUZA, DE J.; HOFFMAN, A. Pipeline Leak Detection and Theft Detection Using Rarefaction Waves. **In 6th Pipeline Technology Conference**, p. 1–12, 2011.

WANG, C.; OLSON, M.; DOIJKHAND, N.; SINGH, S. A novel DdTS technology based on fiber optics for early leak detection in pipelines. **Proceedings - International Carnahan Conference on Security Technology**, p. 1–8, 2017.

ZADKARAMI, M.; SHAHBAZIAN, M.; SALAHSHOOR, K. Pipeline leakage detection and isolation: An integrated approach of statistical and wavelet feature extraction with multi-layer perceptron neural network (MLPNN). **Journal of Loss Prevention in the Process Industries**, v. 43, p. 479–487, 2016.

ZHANG, D.; QIAN, L.; MAO, B.; HUANG, C.; HUANG, B.; SI, Y. A Data-Driven Design for Fault Detection of Wind Turbines Using Random Forests and XGboost. **IEEE Access**, v. 6, p. 21020–21031, 2018.

ZHANG, J.; HOFFMAN, A.; KANE, A.; LEWIS, J. Development of Pipeline Leak Detection Technologies. **Proceedings of 2014 10th International Pipeline Conference**, p. 1–8, 2014.

APÊNDICE A

Hiperparâmetros adotados para os algoritmos de aprendizado de máquina.

Cada uma das tabelas a seguir apresenta os hiperparâmetros de um dos modelos aplicado a regressão e a classificação. A Tabela A - 2 contêm somente os hiperparâmetros para a classificação, uma vez que a regressão logística só se aplica a esse fim.

Tabela A - 1 - Hiperparâmetros adotados para o modelo KNN aplicado a regressão e classificação.

K Vizinhos mais Próximos		
Hiperparâmetro	Classificação	Regressão
n_neighbors	7	11
p	–	2
metric	minkowski	
metric_params	*None*	
n_jobs	-1	
weights	*distance*	
leaf_size	30	
algorithm		

Tabela A - 2 Hiperparâmetros adotados para o modelo Regressão Logística.

Regressão Logística	
Hiperparâmetro	Classificação
C	1000
class_weight	*None*
dual	*False*
fit_intercept	*True*
intercept_scaling	1
max_iter	100
multi_class	*multinomial*
n_jobs	1
penalty	l2
random_state	*None*
solver	lbfgs
tol	0.0001

Tabela A - 3 Hiperparâmetros adotados para o modelo Máquinas de Vetores Suporte (kernel linear e rbf) aplicado a regressão e classificação.

SVM				
Hiperparâmetro	Classificação		Regressão	
C	3190	3450	1	200
random_state	*None*	*None*	–	–
gamma	auto	1	auto	2
kernel	linear	rbf	linear	rbf
verbose	–	–	*False*	*False*
epsilon	–	–	1	0,2
probability	*False*	*False*	–	–
decision_function_shape	ovr	ovr	–	–
class_weight	*None*	*None*	–	–
cache_size	200			
coef0	0			
degree	3			
max_iter	-1			
shrinking	*True*			
tol	0,001			

Tabela A - 4 - Hiperparâmetros adotados para o modelo Redes Neurais aplicado a regressão e classificação.

Rede Neural		
Hiperparâmetro	Classificação	Regressão
hidden_layer_sizes	(26,26,26)	(76,76,76)
activation	relu	
alpha	0,00001	
batch_size	auto	
beta_1	0,9	
beta_2	0,999	
early_stopping	*False*	
epsilon	0,00000001	
learning_rate	*constant*	
learning_rate_init	0,001	
max_iter	5000	
momentum	0,9	
nesterovs_momentum	*True*	
power_t	0,5	
random_state	1	
shuffle	*True*	
solver	lbfgs	
tol	0,0001	
validation_fraction	0,1	
verbose	*False*	
warm_start	*False*	

Tabela A - 5 - Hiperparâmetros adotados para o modelo Floresta Aleatória aplicado a classificação e regressão.

Floresta Aleatória		
Hiperparâmetro	Classificação	Regresão
class_weight	*None*	–
criterion	gini	mse
max_depth	50	*None*
max_features	sqrt	auto
min_impurity_split	*None*	0
min_samples_leaf	1	*None*
min_samples_split	2	1
min_weight_fraction_leaf	0	2
n_estimators	1400	1500
oob_score	*False*	1
random_state	*None*	*False*
verbose	0	*None*
warm_start	*False*	–
max_leaf_nodes	*None*	
min_impurity_decrease	0	
n_jobs	1	
bootstrap	*True*	

Tabela A - 6 - Hiperparâmetros adotados para o modelo Adaboost aplicado a classificação e regressão.

Adaboost		
Hiperparâmetro	Classificação	Regressão
base_estimator	*DecisionTreeClassifier*	*DecisionTreeRegressor*
algorithm	*SAMME*	–
class_weight	*None*	–
criterion	entropy	mse
max_depth	7	12
min_samples_leaf	4	1
n_estimators	1500	1400
random_state	*None*	
max_features	*None*	
max_leaf_nodes	*None*	
min_impurity_decrease	0	
min_impurity_split	*None*	
min_samples_split	2	
min_weight_fraction_leaf	0	
presort	*False*	
random_state	*None*	
splitter	*best*	
learning_rate	1	

Tabela A - 7 – Hiperparâmetros adotados para o modelo Xgboost aplicado a classificação e regressão.

Xgboost		
Hiperparâmetro	Classificação	Regressão
base_estimator	*DecisionTreeClassifier*	*DecisionTreeRegressor*
base_score	0,5	0.5
booster	gbtree	gbtree
learning_rate	0,1	0.08
max_depth	6	7
n_estimators	600	500
objective	*binary logistic*	*reg:linear*
subsample	1	0.75
colsample_bylevel	1	
colsample_bytree	1	
gamma	0	
max_delta_step	0	
min_child_weight	1	
missing	*None*	
n_jobs	1	
nthread	*None*	
random_state	0	
reg_alpha	0	
reg_lambda	1	
scale_pos_weight	1	
seed	*None*	
silent	*True*	

Apêndice B

Resultados completos obtidos com os algoritmos de classificação.

As cinco primeiras tabelas contêm os resultados obtidos com os algoritmos de classificação quando dados de todos os experimentos foram apresentados aos modelos na etapa de treinamento. Na primeira coluna dessas tabelas estão os resultados da validação cruzada, cada um desses valores é a média das acurácias dos cinco treinamentos acompanhado do desvio padrão. Já nas cinco últimas tabelas estão os resultados da classificação de experimentos que não estavam presentes no treinamento dos modelos.

Tabela B - 1 - Resultados obtidos com os algoritmos de classificação quando dados de todos os experimentos foram apresentados aos modelos no treinamento (microfone 1).

Microfone 1				
Modelo	Acurácia Validação Cruzada	Acurácia	Precisão	Sensibilidade
KNN[a]	0,215 ± 0,022	0,220	0,986	0,810
Regressão Logística[a]	0,739 ± 0,040	0,765	0,996	0,958
Rede Neural[a]	0,740 ± 0,014	0,607	0,991	0,829
SVM – Linear[a]	0,839 ± 0,012	0,833	0,974	0,996
SVM – RBF[a]	0,734 ± 0,014	0,753	0,949	0,996
Floresta Aleatória[a]	0,894 ± 0,016	0,884	1,000	1,000
AdaBoost[a]	0,900 ± 0,019	0,872	0,985	1,000
XgBoost[a]	0,901 ± 0,009	0,878	1,000	1,000
Floresta Aleatória[b]	0,995 ± 0,017	0,997	1,000	1,000
Adaboost[b]	0,998 ± 0,008	1,000	1,000	1,000
Xgboost[b]	0,975 ± 0,012	0,982	1,000	1,000

[a] – Modelos alimentados com os dados contendo os componentes principais.
[b] – Modelos alimentados com os dados contendo os atributos mais importantes.

Tabela B - 2 - Resultados obtidos com os algoritmos de classificação quando dados de todos os experimentos foram apresentados aos modelos no treinamento (microfone 2).

	Microfone 2			
Modelo	Acurácia Validação Cruzada	Acurácia	Precisão	Sensibilidade
KNN[a]	0,221 ± 0,012	0,241	0,986	0,821
Regressão Logística[a]	0,782 ± 0,022	0,753	0,988	0,962
Rede Neural[a]	0,760 ± 0,037	0,682	0,995	0,711
SVM – Linear[a]	0,889 ± 0,1387	0,872	0,978	1,000
SVM – RBF[a]	0,825 ± 0,019	0,833	0,974	1,000
Floresta Aleatória[a]	0,885 ± 0,014	0,881	0,992	1,000
AdaBoost[a]	0,904 ± 0,009	0,896	0,989	1,000
XgBoost[a]	0,910 ± 0,017	0,914	0,996	1,000
Floresta Aleatória[b]	0,911 ± 0,011	0,920	0,963	1,000
Adaboost[b]	0,957 ± 0,016	0,964	0,981	1,000
Xgboost[b]	0,866 ± 0,013	0,899	0,985	0,996

[a] – Modelos alimentados com os dados contendo os componentes principais.
[b] – Modelos alimentados com os dados contendo os atributos mais importantes.

Tabela B - 3 - Resultados obtidos com os algoritmos de classificação quando dados de todos os experimentos foram apresentados aos modelos no treinamento (microfone 3).

	Microfone 3			
Modelo	Acurácia Validação Cruzada	Acurácia	Precisão	Sensibilidade
KNN[a]	0,204 ± 0,022	0,223	0,968	0,798
Regressão Logística[a]	0,674 ± 0,015	0,583	0,889	1,000
Rede Neural[a]	0,610 ± 0,032	0,616	0,984	0,692
SVM – Linear[a]	0,774 ± 0,030	0,753	0,996	1,000
SVM – RBF[a]	0,784 ± 0,034	0,780	0,985	1,000
Floresta Aleatória[a]	0,800 ± 0,025	0,792	0,985	1,000
AdaBoost[a]	0,799 ± 0,022	0,801	0,989	1,000
XgBoost[a]	0,819 ± 0,015	0,789	0,992	0,996
Floresta Aleatória[b]	0,977 ± 0,008	0,985	0,989	1,000
Adaboost[b]	0,993 ± 0,032	1,000	1,000	1,000
Xgboost[b]	0,943 ± 0,027	0,961	0,996	1,000

[a] – Modelos alimentados com os dados contendo os componentes principais.
[b] – Modelos alimentados com os dados contendo os atributos mais importantes.

Tabela B - 4 - Resultados obtidos com os algoritmos de classificação quando dados de todos os experimentos foram apresentados aos modelos no treinamento (microfone 4).

Microfone 4				
Modelo	Acurácia Validação Cruzada	Acurácia	Precisão	Sensibilidade
KNN[a]	0,215 ± 0,015	0,205	0,963	0,798
Regressão Logística[a]	0,602 ± 0,027	0,673	0,969	0,951
Rede Neural[a]	0,587 ± 0,033	0,589	1,000	0,643
SVM – Linear[a]	0,747 ± 0,018	0,756	0,970	0,996
SVM – RBF[a]	0,757 ± 0,018	0,744	0,953	0,996
Floresta Aleatória[a]	0,775 ± 0,014	0,753	0,989	0,996
AdaBoost[a]	0,772 ± 0,023	0,804	0,996	1,000
XgBoost[a]	0,776 ± 0,027	0,813	0,996	0,992
Floresta Aleatória[b]	0,921 ± 0,021	0,932	0,978	0,996
Adaboost[b]	0,934 ± 0,023	0,940	0,985	1,000
Xgboost[b]	0,917 ± 0,012	0,923	0,989	0,992

[a] – Modelos alimentados com os dados contendo os componentes principais.
[b] – Modelos alimentados com os dados contendo os atributos mais importantes.

Tabela B - 5 - Resultados obtidos com os algoritmos de classificação quando dados de todos os experimentos foram apresentados aos modelos no treinamento (microfone 5).

Microfone 5				
Modelo	Acurácia Validação Cruzada	Acurácia	Precisão	Sensibilidade
KNN[a]	0,228 ± 0,017	0,235	0,968	0,814
Regressão Logística[a]	0,599 ± 0,023	0,557	0,967	0,776
Rede Neural[a]	0,457 ± 0,042	0,354	1,000	0,433
SVM – Linear[a]	0,851 ± 0,021	0,863	0,974	0,992
SVM – RBF[a]	0,862 ± 0,017	0,851	0,963	0,992
Floresta Aleatória[a]	0,817 ± 0,011	0,821	1,000	0,996
AdaBoost[a]	0,844 ± 0,014	0,839	0,989	0,996
XgBoost[a]	0,855 ± 0,009	0,842	0,992	0,996
Floresta Aleatória[b]	0,952 ± 0,008	0,961	0,996	1,00
Adaboost[b]	0,961 ± 0,013	0,970	0,996	1,00
Xgboost[b]	0,944 ± 0,012	0,958	0,992	1,00

[a] – Modelos alimentados com os dados contendo os componentes principais.
[b] – Modelos alimentados com os dados contendo os atributos mais importantes.

Tabela B - 6 - Resultados da classificação de experimentos cujos dados não foram apresentados aos modelos no treinamento (microfone 1).

Microfone 1							
	Experimento fora do treinamento						
Modelo	3	4	5	6	7	8	9
KNN[a]	0,98	0,15	0,78	1,00	0,54	0,48	0,13
Regressão Logística[a]	0,88	0,96	0,98	1,00	1,00	1,00	0,96
Rede Neural[a]	0,98	0,98	0,00	1,00	1,00	1,00	0,98
SVM – Linear[a]	0,99	0,99	1,00	1,00	1,00	1,00	0,99
SVM – RBF[a]	0,99	0,99	1,00	1,00	1,00	1,00	0,99
Floresta Aleatória[a]	1,00	1,00	1,00	1,00	1,00	1,00	1,00
AdaBoost[a]	1,00	1,00	0,01	1,00	0,00	0,81	1,00
XgBoost[a]	0,98	1,00	0,93	0,99	1,00	1,00	1,00
Floresta Aleatória[b]	1,00	1,00	1,00	1,00	1,00	1,00	1,00
Adaboost[b]	1,00	1,00	1,00	1,00	1,00	1,00	1,00
Xgboost[b]	0,97	0,98	1,00	1,00	1,00	1,00	1,00

[a] – Modelos alimentados com os dados contendo os componentes principais.
[b] – Modelos alimentados com os dados contendo os atributos mais importantes.

Tabela B - 7 - Resultados da classificação de experimentos cujos dados não foram apresentados aos modelos no treinamento (microfone 2).

Microfone 2							
	Experimento fora do treinamento						
Modelo	3	4	5	6	7	8	9
KNN[a]	1,00	0,18	0,79	0,99	0,13	0,49	0,21
Regressão Logística[a]	0,99	0,98	0,93	0,95	1,00	1,00	0,99
Rede Neural[a]	1,00	1,00	0,00	0,99	1,00	1,00	0,95
SVM – Linear[a]	1,00	0,98	1,00	1,00	1,00	1,00	1,00
SVM – RBF[a]	1,00	0,99	0,59	1,00	1,00	1,00	1,00
Floresta Aleatória[a]	1,00	1,00	0,99	1,00	1,00	1,00	1,00
AdaBoost[a]	1,00	1,00	0,99	1,00	0,00	1,00	1,00
XgBoost[a]	1,00	1,00	0,99	1,00	1,00	1,00	1,00
Floresta Aleatóriab	0,16	1,00	0,85	1,00	1,00	0,99	1,00
Adaboost[b]	0,48	1,00	0,87	1,00	1,00	1,00	1,00
Xgboost[b]	0,53	0,99	0,93	1,00	1,00	0,99	1,00

[a] – Modelos alimentados com os dados contendo os componentes principais.
[b] – Modelos alimentados com os dados contendo os atributos mais importantes.

Tabela B - 8 - Resultados da classificação de experimentos cujos dados não foram apresentados aos modelos no treinamento (microfone 3).

Microfone 3							
	Experimento fora do treinamento						
Model	3	4	5	6	7	8	9
KNN[a]	1,00	0,042	0,83	1,00	0,92	0,99	0,06
Regressão Logística[a]	0,99	0,62	1,00	1,00	1,00	0,92	0,90
Rede Neural[a]	1,00	0,96	0,00	0,99	1,00	0,01	0,00
SVM – Linear[a]	1,00	0,98	1,00	1,00	1,00	1,00	1,00
SVM – RBF[a]	1,00	0,99	1,00	1,00	1,00	0,78	1,00
Floresta Aleatória[a]	1,00	1,00	0,67	1,00	0,99	0,95	1,00
AdaBoost[a]	1,00	1,00	0,00	1,00	0,76	0,69	1,00
XgBoost[a]	0,99	0,97	0,13	1,00	0,83	0,65	1,00
Floresta Aleatória[b]	1,00	1,00	0,08	1,00	0,08	0,73	1,00
Adaboost[b]	1,00	1,00	0,02	1,00	1,00	0,98	1,00
Xgboost[b]	1,00	0,98	0,56	0,98	0,99	1,00	1,00

[a] – Modelos alimentados com os dados contendo os componentes principais.
[b] – Modelos alimentados com os dados contendo os atributos mais importantes.

Tabela B - 9 -- Resultados da classificação de experimentos cujos dados não foram apresentados aos modelos no treinamento (microfone 4).

Microfone 4							
	Experimento fora do Treinamento						
Modelo	3	4	5	6	7	8	9
KNN[a]	0,99	0,10	0,92	1,00	0,99	1,00	0,03
Regressão Logística[a]	1,00	0,00	0,97	1,00	0,89	0,99	0,00
Rede Neural[a]	1,00	0,77	0,00	0,99	0,34	0,00	0,78
SVM – Linear[a]	1,00	0,97	1,00	1,00	1,00	1,00	0,96
SVM – RBF[a]	1,00	0,95	1,00	1,00	0,00	1,00	0,94
Floresta Aleatória[a]	1,00	0,98	0,82	1,00	0,65	0,92	0,98
AdaBoost[a]	1,00	0,98	0,64	1,00	0,65	0,99	0,98
XgBoost[a]	0,99	0,96	0,97	1,00	0,65	0,99	0,98
Floresta Aleatória[b]	1,00	0,94	0,02	1,00	0,00	0,68	1,00
Adaboost[b]	1,00	0,92	0,01	1,00	0,08	0,66	0,98
Xgboost[b]	1,00	0,82	0,46	0,99	0,23	0,98	0,94

[a] – Modelos alimentados com os dados contendo os componentes principais.
[b] – Modelos alimentados com os dados contendo os atributos mais importantes.

Tabela B - 10 - Resultados da classificação de experimentos cujos dados não foram apresentados aos modelos no treinamento (microfone 5).

	Microfone 5						
	Experimento fora do Treinamento						
Model	3	4	5	6	7	8	9
KNN[a]	0,99	0,06	0,86	0,98	0,99	1,00	0,07
Regressão Logística[a]	0,99	0,00	0,98	1,00	0,93	1,00	0,13
Rede Neural[a]	0,81	0,18	1,00	1,00	0,78	0,00	0,90
SVM – Linear[a]	1,00	0,82	1,00	1,00	1,00	1,00	0,82
SVM – RBF[a]	1,00	0,80	1,00	1,00	0,01	1,00	0,88
Floresta Aleatória[a]	1,00	0,99	0,97	1,00	0,96	1,00	0,99
AdaBoost[a]	0,97	0,97	0,95	1,00	0,99	1,00	0,98
XgBoost[a]	0,99	0,98	0,74	0,99	0,98	0,96	0,98
Floresta Aleatória[b]	1,00	0,48	0,16	1,00	0,00	0,69	1,00
Adaboost[b]	1,00	0,46	0,03	0,98	0,18	0,75	1,00
Xgboost[b]	1,00	0,46	0,55	0,92	0,21	0,85	1,00

[a] – Modelos alimentados com os dados contendo os componentes principais.
[b] – Modelos alimentados com os dados contendo os atributos mais importantes.

Apêndice C

Resultados completos obtidos com os algoritmos de regressão.

Nas cinco tabelas iniciais estão os resultados obtidos (RMSE) com os algoritmos de regressão na validação cruzada e com os dados separados para validação dos modelos. Cada resultado da validação cruzada é a média dos cinco erros quadráticos médios obtidos nos cinco treinamentos acompanhada do desvio padrão. As tabelas seguintes apresentam as predições dos modelos quando dados de todos os experimentos foram apresentados aos modelos no treinamento e as cinco finais as predições para os experimentos ausentes do treinamento dos modelos

Tabela C - 1 - Resultados obtidos com os algoritmos de regressão quando dados de todos os experimentos foram apresentados aos modelos (microfone 1).

Microfone 1		
Modelo	RMSE Validação Cruzada (m)	RMSE Validação (m)
KNN[a]	2,40 ± 0,094	2,23
Rede Neural[a]	20,4 ± 26,2	1,23
SVM – Linear[a]	1,72 ± 0,13	1,72
SVM – RBF[a]	1,73 ± 0,066	1,65
Floresta Aleatória[a]	1,35 ± 0,101	1,29
AdaBoost[a]	1,67 ± 0,197	1,62
XgBoost[a]	1,39 ± 0,110	1,28
Floresta Aleatória[b]	0.78 ± 0,076	0,60
Adaboost[b]	0.29 ± 0.12	0,25
Xgboost[b]	0.65 ±0.091	0,56

[a] – Modelos alimentados com os dados contendo os componentes principais.
[b] – Modelos alimentados com os dados contendo os atributos mais importantes.

Tabela C - 2 - Resultados obtidos com os algoritmos de regressão quando dados de todos os experimentos foram apresentados aos modelos (microfone 2).

Microfone 2		
Modelo	RMSE Validação Cruzada (m)	RMSE Validação (m)
K Vizinhos mais Próximos[a]	2,36 ± 0,11	2,23
Rede Neural[a]	3,45 ± 3,97	1,41
SVM – Linear[a]	1,71 ± 0,16	1,74
SVM – RBF[a]	1,53 ± 0,092	1,55
Floresta Aleatória[a]	1,29 ± 0,073	1,23
AdaBoost[a]	1,60 ± 0,15	1,35
XgBoost[a]	1,33 ± 0,062	1,17
Floresta Aleatória[b]	0,72 ± 0,012	0,68
Adaboost[b]	0,23 ± 0,014	0,00
Xgboost[b]	0,88 ± 0.143	0,63

[a] – Modelos alimentados com os dados contendo os componentes principais.
[b] – Modelos alimentados com os dados contendo os atributos mais importantes.

Tabela C - 3 - Resultados obtidos com os algoritmos de regressão quando dados de todos os experimentos foram apresentados aos modelos (microfone 3).

Microfone 3		
Modelo	RMSE Validação Cruzada (m)	RMSE Validação (m)
K Vizinhos mais Próximos[a]	2,20 ± 0,083	2,11
Rede Neural[a]	1,42 ± 0,069	1,40
SVM – Linear[a]	1,70 ± 0,093	1,77
SVM – RBF[a]	1,73 ± 0,087	1,75
Floresta Aleatória[a]	1,42 ± 0,086	1,42
AdaBoost[a]	1,94 ± 0,068	2,02
XgBoost[a]	1,46 ± 0,067	1,54
Floresta Aleatória[b]	1,01 ± 0,054	0,96
Adaboost[b]	0,76 ± 0,087	0,54
Xgboost[b]	0,97 ± 0.065	0,89

[a] – Modelos alimentados com os dados contendo os componentes principais.
[b] – Modelos alimentados com os dados contendo os atributos mais importantes.

Tabela C - 4 - Resultados obtidos com os algoritmos de regressão quando dados de todos os experimentos foram apresentados aos modelos (microfone 4).

Microfone 4		
Modelo	RMSE Validação Cruzada (m)	RMSE Validação (m)
K Vizinhos mais Próximos[a]	1,99± 0,11	1,98
Rede Neural[a]	1,42 ± 0,078	1,37
SVM – Linear[a]	1,69 ± 0,088	1,77
SVM – RBF[a]	1,78 ± 0,077	1,77
Floresta Aleatória[a]	1,39 ±0,099	1,36
AdaBoost[a]	1,69 ± 0,13	1,80
XgBoost[a]	1,45± 0,104	1,41
Floresta Aleatória[b]	0,74 ± 0,087	0,65
Adaboost[b]	0,34 ± 0,054	0,29
Xgboost[b]	0,76 ± 0,098	0,58

[a] – Modelos alimentados com os dados contendo os componentes principais.
[b] – Modelos alimentados com os dados contendo os atributos mais importantes.

Tabela C - 5 - Resultados obtidos com os algoritmos de regressão quando dados de todos os experimentos foram apresentados aos modelos (microfone 5).

Microfone 5		
Modelo	RMSE Validação Cruzada (m)	RMSE Validação (m)
K Vizinhos mais Próximos[a]	1,85 ± 0,10	1,845
Rede Neural[a]	1,28 ± 0,115	1,41
SVM – Linear[a]	1,72 ± 0,090	1,78
SVM – RBF[a]	1,65 ± 0,0769	1,67
Floresta Aleatória[a]	1,19 ± 0,126	1,21
AdaBoost[a]	1,27 ± 0,171	1,32
XgBoost[a]	1,17 ± 0,118	1,21
Floresta Aleatória[b]	0,474 ± 0,113	0,35
Adaboost[b]	0,238 ± 0,08	0,14
Xgboost[b]	0,451 ± 0,06	0,36

[a] – Modelos alimentados com os dados contendo os componentes principais.
[b] – Modelos alimentados com os dados contendo os atributos mais importantes.

Tabela C - 6 - Predições quando dados de todos os experimentos estavam presentes na etapa de treinamento dos modelos (microfone 1).

Microfone 1														
	VR = 0,2 m		VR = 1,0 m		VR = 1,6 m		VR = 2,4 m		VR = 3,4 m		VR = 4,8 m		VR = 5,6 m	
Modelo	VP (m)	EL (%)	VP (m)	EL (%)	VP (m)	EL (%)	VP (m)	EL (%)	VP (m)	EL (%)	VP (m)	EL (%)	VP (m)	EL (%)
KNN[a]	2,81	5,49	3,13	4,48	1,12	1,02	3,25	1,79	1,64	3,70	1,23	7,51	3,29	4,86
NN[a]	0,68	1,02	2,62	3,41	1,55	0,10	1,63	1,63	3,62	0,46	4,40	0,85	3,37	4,70
SVM – Linear[a]	1,68	3,13	1,97	2,05	1,33	0,57	1,66	1,56	4,24	1,76	3,61	2,50	1,97	7,64
SVM – RBF[a]	0,44	0,50	1,93	1,95	3,12	3,20	1,88	1,10	3,22	0,37	3,22	3,32	2,09	7,38
Floresta Aleatória[a]	1,23	2,18	3,14	4,51	1,56	0,09	1,60	1,69	3,39	0,01	4,64	0,35	3,27	4,90
AdaBoost[a]	0,86	1,39	3,11	4,44	1,60	0,00	1,87	1,12	3,40	0,00	4,80	0,00	3,03	5,40
XgBoost[a]	1,12	1,94	3,19	4,60	1,56	0,09	1,60	1,68	3,40	0,01	4,57	0,49	3,29	4,87
Floresta Aleatória[b]	0,52	0,67	1,70	1,48	1,95	0,75	2,23	0,37	3,43	0,07	4,58	0,46	4,28	2,77
Adaboost[b]	0,2	0,00	1,00	0,00	1,6	0,00	2,4	0,00	3,4	0,00	4,8	0,00	5,6	0,00
Xgboost[b]	0,41	0,45	1,50	1,05	1,90	0,63	2,20	0,43	3,41	0,03	4,58	0,47	4,35	2,64

[a] – Modelos alimentados com os dados contendo os componentes principais.
[b] – Modelos alimentados com os dados contendo os atributos mais importantes.
VR – Posição do orifício.
VP – Valor predito pelo modelo.
EL - Erro de localização.

Tabela C - 7 - Predições quando dados de todos os experimentos estavam presentes na etapa de treinamento dos modelos (microfone 2).

Microfone 2														
	VR = 0,2 m		VR = 1,0 m		VR = 1,6 m		VR = 2,4 m		VR = 3,4 m		VR = 4,8 m		VR = 5,6 m	
Modelo	VP (m)	EL (%)	VP (m)	EL (%)	VP (m)	EL (%)	VP (m)	EL (%)	VP (m)	EL (%)	VP (m)	EL (%)	VP (m)	EL (%)
KNN[a]	3,29	6,50	2,93	4,07	1,23	2,45	3,35	1,99	1,50	3,99	1,14	7,70	3,12	5,23
NN[a]	0,71	1,08	2,76	3,71	1,64	1,59	1,60	1,69	3,40	0,00	4,47	0,70	3,58	4,25
SVM – Linear[a]	1,89	3,55	1,99	2,08	2,02	0,79	1,89	1,08	4,00	1,27	3,89	1,91	1,99	7,60
SVM – RBF[a]	0,57	0,78	1,40	0,85	2,12	0,59	1,99	0,86	3,71	0,65	3,71	2,30	1,45	8,74
Floresta Aleatória[a]	1,08	1,86	3,13	4,48	1,59	1,71	1,62	1,65	3,42	0,05	4,70	0,20	3,46	4,50
AdaBoost[a]	0,86	1,39	2,37	2,88	1,60	1,68	2,03	0,77	3,40	0,00	4,80	0,00	4,23	2,89
XgBoost[a]	0,99	1,66	3,03	4,27	1,56	1,78	1,57	1,75	3,43	0,06	4,69	0,24	3,60	4,22
Floresta Aleatória[b]	0,48	0,60	1,85	1,80	1,61	1,67	2,39	0,02	3,48	0,16	4,65	0,31	4,11	3,14
Adaboost[b]	0,20	0,00	1,00	0,00	1,6	1,68	2,40	0,00	3,40	0,00	4,80	0,00	5,60	0,00
Xgboost[b]	0,38	0,37	1,67	1,40	1,59	1,70	2,41	0,01	3,46	0,12	4,76	0,08	4,34	2,65

[a] – Modelos alimentados com os dados contendo os componentes principais.
[b] – Modelos alimentados com os dados contendo os atributos mais importantes.
VR – Posição do orifício.
VP – Valor predito pelo modelo.
EL - Erro de localização.

Tabela C - 8 - Predições quando dados de todos os experimentos estavam presentes na etapa de treinamento dos modelos (microfone 3).

Microfone 3														
	VR = 0,2 m		VR = 1,0 m		VR = 1,6 m		VR = 2,4 m		VR = 3,4 m		VR = 4,8 m		VR = 5,6 m	
Modelo	VP (m)	EL (%)	VP (m)	EL (%)	VP (m)	EL (%)	VP (m)	EL (%)	VP (m)	EL (%)	VP (m)	EL (%)	VP (m)	EL (%)
KNN[a]	3,23	6,38	3,38	5,01	1,39	0,45	3,27	1,84	1,98	2,98	1,60	6,74	2,99	5,49
NN[a]	1,28	2,27	3,20	4,63	1,77	0,35	1,30	2,31	3,43	0,06	4,65	0,32	3,41	4,60
SVM – Linear[a]	2,05	3,89	1,99	2,09	2,54	1,99	2,05	0,73	4,06	1,39	3,86	1,99	1,99	7,59
SVM – RBF[a]	1,17	2,05	1,47	0,99	2,76	2,44	1,90	1,05	3,87	0,99	4,22	1,21	1,50	8,64
Floresta Aleatória[a]	1,33	2,39	3,20	4,64	1,55	0,10	1,43	2,04	3,40	0,00	4,80	0,00	3,30	4,85
AdaBoost[a]	1,47	2,67	2,92	4,04	1,60	0,00	2,02	0,79	3,40	0,00	4,80	0,00	3,10	5,25
XgBoost[a]	1,36	2,44	3,14	4,50	1,61	0,03	1,46	1,99	3,42	0,05	4,76	0,09	3,43	4,57
Floresta Aleatória[b]	0,93	1,54	2,21	2,55	1,88	0,60	1,78	1,31	3,51	0,24	4,40	0,84	3,79	3,80
Adaboost[b]	0,325	0,26	1,25	0,53	1,6	0,00	2,4	0,00	3,4	0,00	4,8	0,00	5,067	1,12
Xgboost[b]	0,79	1,24	2,22	2,57	1,82	0,47	1,79	1,28	3,49	0,20	4,41	0,82	4,10	3,16

[a] – Modelos alimentados com os dados contendo os componentes principais.
[b] – Modelos alimentados com os dados contendo os atributos mais importantes.
VR – Posição do orifício.
VP – Valor predito pelo modelo.
EL - Erro de localização

Tabela C - 9 - Predições quando dados de todos os experimentos estavam presentes na etapa de treinamento dos modelos (microfone 4).

Microfone 4														
	VR = 0,2 m		VR = 1,0 m		VR = 1,6 m		VR = 2,4 m		VR = 3,4 m		VR = 4,8 m		VR = 5,6 m	
Modelo	VP (m)	EL (%)	VP (m)	EL (%)	VP (m)	EL (%)	VP (m)	EL (%)	VP (m)	EL (%)	VP (m)	EL (%)	VP (m)	EL (%)
KNN[a]	2,82	5,51	3,45	5,15	1,47	0,27	2,37	0,06	4,04	1,35	1,93	6,05	3,31	4,83
NN[a]	1,29	2,30	3,26	4,77	1,61	0,02	1,32	2,28	3,49	0,19	4,77	0,05	3,33	4,77
SVM – Linear[a]	2,05	3,90	1,99	2,09	2,54	1,98	2,06	0,71	3,80	0,85	3,75	2,21	1,99	7,60
SVM – RBF[a]	1,30	2,31	1,29	0,62	1,80	0,42	1,33	2,25	3,82	0,87	4,71	0,18	1,33	8,98
Floresta Aleatória[a]	1,28	2,27	3,21	4,65	1,56	0,08	1,45	2,00	3,40	0,00	4,80	0,00	3,11	5,24
AdaBoost[a]	1,39	2,51	2,78	3,75	1,63	0,07	1,87	1,11	3,40	0,00	4,80	0,00	2,52	6,49
XgBoost[a]	1,28	2,27	3,07	4,36	1,60	0,01	1,50	1,90	3,40	0,00	4,76	0,09	3,19	5,07
Floresta Aleatória[b]	0,81	1,29	1,42	0,89	1,77	0,37	2,18	0,47	3,48	0,16	4,49	0,65	4,72	1,86
Adaboost[b]	0,47	0,56	1,06	0,12	1,6	0,00	2,28	0,39	3,4	0,00	4,8	0,00	5,6	0,00
Xgboost[b]	0,70	1,06	1,58	1,21	1,79	0,41	2,17	0,49	3,45	0,11	4,53	0,56	4,82	1,65

[a] – Modelos alimentados com os dados contendo os componentes principais.
[b] – Modelos alimentados com os dados contendo os atributos mais importantes.
VR – Posição do orifício.
VP – Valor predito pelo modelo.
EL - Erro de localização.

Tabela C - 10 - Predições quando dados de todos os experimentos estavam presentes na etapa de treinamento dos modelos (microfone 5).

Microfone 5														
	VR = 0,2 m		VR = 1,0 m		VR = 1,6 m		VR = 2,4 m		VR = 3,4 m		VR = 4,8 m		VR = 5,6 m	
Modelo	VP (m)	EL (%)	VP (m)	EL (%)	VP (m)	EL (%)	VP (m)	EL (%)	VP (m)	EL (%)	VP (m)	EL (%)	VP (m)	EL (%)
KNN[a]	2,34	4,50	3,21	4,65	1,48	0,25	2,43	0,06	4,61	2,54	2,06	5,76	4,25	2,84
NN[a]	0,94	1,55	1,74	1,56	1,57	0,07	1,94	0,98	3,42	0,03	4,79	0,03	4,98	1,31
SVM – Linear[a]	2,05	3,90	1,99	2,09	2,53	1,97	2,06	0,72	4,09	1,44	3,68	2,35	2,00	7,59
SVM – RBF[a]	0,82	1,31	1,38	0,81	1,10	1,05	1,79	1,28	3,68	0,58	4,52	0,60	1,77	8,07
Floresta Aleatória[a]	1,07	1,83	2,86	3,91	1,58	0,04	1,50	1,89	3,40	0,00	4,80	0,00	3,80	3,79
AdaBoost[a]	1,14	1,97	1,82	1,74	1,60	0,00	2,01	0,82	3,40	0,00	4,80	0,00	4,24	2,87
XgBoost[a]	1,00	1,68	2,56	3,28	1,62	0,05	1,44	2,02	3,38	0,04	4,78	0,04	4,01	3,34
Floresta Aleatória[b]	0,32	0,26	1,23	0,49	1,96	0,76	2,34	0,12	3,52	0,25	4,51	0,60	5,17	0,90
Adaboost[b]	0,27	0,14	1,06	0,12	1,60	0,00	2,40	0,00	3,40	0,00	4,80	0,00	5,41	0,40
Xgboost[b]	0,31	0,23	1,27	0,57	1,85	0,52	2,37	0,06	3,51	0,22	4,57	0,48	5,22	0,79

[a] – Modelos alimentados com os dados contendo os componentes principais.
[b] – Modelos alimentados com os dados contendo os atributos mais importantes.
VR – Posição do orifício.
VP – Valor predito pelo modelo.
EL - Erro de localização.

Tabela C - 11 - Predições para os experimentos cujos dados não foram apresentados os modelos na etapa de treinamento (microfone 1).

Microfone 1										
	VR = 1,0 m		VR = 1,6 m		VR = 2,4 m		VR = 3,4 m		VR = 4,8 m	
Modelo	VP (m)	EL (%)	VP (m)	EL (%)	VP (m)	EL (%)	VP (m)	EL (%)	VP (m)	EL (%)
KNN[a]	5,6	9,68	1,11	1,03	3,19	1,66	3,26	0,29	1,23	7,52
NN[a]	5,5	9,47	-16,47	38,04	0,3	4,42	114,79	234,51	3,35	3,05
SVM – Linear[a]	3,67	5,62	0,62	2,06	1,67	1,54	7,31	8,23	1,69	6,55
SVM – RBF[a]	5,49	9,45	3,87	4,77	3,38	2,05	3,14	0,55	2,42	5,00
Floresta Aleatória[a]	5,59	9,67	3,67	4,35	0,36	4,28	3,80	0,84	1,52	6,90
AdaBoost[a]	5,60	9,68	4,01	5,07	0,24	4,55	4,80	2,95	1,60	6,74
XgBoost[a]	5,59	9,66	3,40	3,78	0,32	4,38	3,77	0,79	1,43	7,09
Floresta Aleatória[b]	5,28	9,01	4,63	6,38	0,78	3,41	3,09	0,65	2,5	4,84
Adaboost[b]	5,6	9,68	4,8	6,74	0,71	3,56	4,8	2,95	1,6	6,74
Xgboost[b]	0,93	0,15	1,58	0,04	2,1	0,63	1,94	3,07	3,1	3,58

[a] – Modelos alimentados com os dados contendo os componentes principais.
[b] – Modelos alimentados com os dados contendo os atributos mais importantes.
VR – Posição do orifício.
VP – Valor predito pelo modelo.
EL - Erro de localização

Tabela C - 12 - Predições para os experimentos cujos dados não foram apresentados os modelos na etapa de treinamento (microfone 2).

Microfone 2										
	VR = 1,0 m		VR = 1,6 m		VR = 2,4 m		VR = 3,4 m		VR = 4,8 m	
Modelo	VP (m)	EL (%)	VP (m)	EL (%)	VP (m)	EL (%)	VP (m)	EL (%)	VP (m)	EL (%)
KNN[a]	5,6	9,68	1,28	0,67	3,36	2,02	1,52	3,96	1,38	7,20
NN[a]	5,7	9,89	2,19	1,24	0,25	4,53	37,4	71,58	3,08	3,62
SVM – Linear[a]	3,23	4,69	1,97	0,78	1,87	1,12	6,00	5,47	1,33	7,31
SVM – RBF[a]	5,57	9,63	3,09	3,13	0,61	3,76	3,64	0,50	2,66	4,51
Floresta Aleatória[a]	5,52	9,52	2,72	2,36	0,39	4,23	4,08	1,42	1,40	7,16
AdaBoost[a]	5,60	9,68	2,40	1,68	0,32	4,38	4,80	2,95	1,97	5,96
XgBoost[a]	5,51	9,50	2,81	2,54	0,36	4,29	4,33	1,95	1,36	7,25
Floresta Aleatória[b]	5,5	9,47	4,36	5,81	1,77	1,33	4,79	2,93	3,16	3,45
Adaboost[b]	5,6	9,68	2,55	2,00	0,93	3,09	4,8	2,95	2,4	5,05
Xgboost[b]	2,1	2,32	1,31	0,61	2,4	0,00	1,76	3,45	2,4	5,05

[a] – Modelos alimentados com os dados contendo os componentes principais.
[b] – Modelos alimentados com os dados contendo os atributos mais importantes.
VR – Posição do orifício.
VP – Valor predito pelo modelo.
EL - Erro de localização

Tabela C - 13 - Predições para os experimentos cujos dados não foram apresentados os modelos na etapa de treinamento (microfone 3).

Microfone 3										
	VR = 1,0 m		VR = 1,6 m		VR = 2,4 m		VR = 3,4 m		VR = 4,8 m	
Modelo	VP (m)	EL (%)	VP (m)	EL (%)	VP (m)	EL (%)	VP (m)	EL (%)	VP (m)	EL (%)
KNN[a]	5,6	9,68	1,35	0,53	3,11	1,49	2,17	2,59	1,6	6,74
NN[a]	5,69	9,87	3,07	3,09	0,29	4,44	6,56	6,65	2,31	5,24
SVM – Linear[a]	1,97	2,04	2,64	2,19	2,03	0,78	6,14	5,77	1,81	6,29
SVM – RBF[a]	5,41	9,27	1,14	0,97	0,25	4,53	4,53	2,38	3,16	3,46
Floresta Aleatória[a]	5,59	9,67	1,43	0,36	0,21	4,61	4,80	2,94	1,81	6,29
AdaBoost[a]	5,60	9,68	2,43	1,74	0,25	4,53	4,80	2,95	2,09	5,70
XgBoost[a]	5,59	9,67	1,60	0,00	0,23	4,57	4,64	2,60	1,73	6,46
Floresta Aleatória[b]	5,22	8,88	4,25	5,58	0,94	3,07	4,79	2,93	2,87	4,06
Adaboost[b]	5,6	9,68	3,95	4,95	0,39	4,23	4,8	2,95	2,03	5,83
Xgboost[b]	1,92	1,94	1,22	0,80	3,6	2,53	1,67	3,64	2,44	4,97

[a] – Modelos alimentados com os dados contendo os componentes principais.
[b] – Modelos alimentados com os dados contendo os atributos mais importantes.
VR – Posição do orifício.
VP – Valor predito pelo modelo.
EL - Erro de localização

Tabela C - 14 - Predições para os experimentos cujos dados não foram apresentados os modelos na etapa de treinamento (microfone 4).

Microfone 4										
	VR = 1,0 m		VR = 1,6 m		VR = 2,4 m		VR = 3,4 m		VR = 4,8 m	
Modelo	VP (m)	EL (%)	VP (m)	EL (%)	VP (m)	EL (%)	VP (m)	EL (%)	VP (m)	EL (%)
KNN[a]	5,6	9,68	1,42	0,38	2,39	0,02	4,42	2,15	1,6	6,74
NN[a]	5,55	9,58	1,44	0,34	0,22	4,59	9,01	11,81	2,48	4,88
SVM – Linear[a]	2,08	2,27	2,61	2,13	2,06	0,72	7,12	7,83	1,93	6,04
SVM – RBF[a]	4,67	7,72	0,80	1,69	0,37	4,27	4,74	2,83	2,60	4,62
Floresta Aleatória[a]	5,60	9,68	1,40	0,42	0,21	4,62	4,80	2,95	1,58	6,78
AdaBoost[a]	5,60	9,68	2,40	1,68	0,35	4,32	4,80	2,95	1,69	6,54
XgBoost[a]	5,60	9,67	1,71	0,22	0,26	4,50	4,71	2,76	1,66	6,62
Floresta Aleatória[b]	5,11	8,65	4,19	5,45	0,66	3,66	4,77	2,88	2,93	3,94
Adaboost[b]	5,57	9,62	3,18	3,33	0,23	4,57	4,8	2,95	2,52	4,80
Xgboost[b]	2,1	2,32	1,99	0,82	1,85	1,16	1,84	3,28	2,72	4,38

[a] – Modelos alimentados com os dados contendo os componentes principais.
[b] – Modelos alimentados com os dados contendo os atributos mais importantes.
VR – Posição do orifício.
VP – Valor predito pelo modelo.
EL - Erro de localização

Tabela C - 15 - Predições para os experimentos cujos dados não foram apresentados os modelos na etapa de treinamento (microfone 5).

Microfone 5										
	VR= 1,0 m		VR= 1,6 m		VR= 2,4 m		VR= 3,4 m		VR= 4,8 m	
Modelo	VP (m)	EL (%)	VP (m)	EL (%)	VP (m)	EL (%)	VP (m)	EL (%)	VP (m)	EL (%)
KNN[a]	5,6	9,68	1,31	0,61	2,14	0,55	4,7	2,74	1,6	6,74
NN[a]	5,6	9,68	3,26	3,49	0,44	4,13	1,06	4,93	2,15	5,58
SVM – Linear[a]	2,16	2,44	2,65	2,21	2,06	0,72	6,97	7,52	2,03	5,83
SVM – RBF[a]	4,69	7,77	0,43	2,47	0,63	3,72	4,75	2,85	3,67	2,38
Floresta Aleatória[a]	5,57	9,61	2,25	1,38	0,29	4,43	4,79	2,92	1,90	6,10
AdaBoost[a]	5,60	9,68	3,39	3,77	0,29	4,44	4,80	2,95	1,79	6,33
XgBoost[a]	5,57	9,61	2,30	1,48	0,24	4,55	4,71	2,75	1,71	6,50
Floresta Aleatória[b]	4,93	8,27	4	5,05	1,05	2,84	4,77	2,88	2,9	4,00
Adaboost[b]	5,55	9,58	4	5,05	0,68	3,62	4,8	2,95	2,56	4,72
Xgboost[b]	0,91	0,19	1,61	0,02	1,9	1,05	1,91	3,14	2,75	4,32

[a] – Modelos alimentados com os dados contendo os componentes principais.
[b] – Modelos alimentados com os dados contendo os atributos mais importantes.
VR – Posição do orifício.
VP – Valor predito pelo modelo.
EL - Erro de localização