Antonio Hevertton Martins Silva

Detecção de anomalia no acionamento de válvulas gavetas submarinas utilizando técnicas de aprendizagem de máquina

Monografia de Final de Curso

03/11/2020

Monografia apresentada ao Departamento de Engenharia Elétrica da PUC/Rio como parte dos requisitos para a obtenção do título de Especialização em Business Intelligence.

Orientadores:

Prof. Dr. Leonardo Alfredo Forero Mendoza

Agradecimentos

À Petrobras pelo suporte financeiro.

À minha esposa, Aline, pelo apoio, incentivo e compreensão.

Aos meus pais, Antonio e Conceição, e irmão que me apoiam em todos os momentos da minha jornada.

Aos professores do BI Master pelo excelente e atualizado conteúdo de todas as disciplinas.

RESUMO

A detecção de anomalias tem aplicação há bastante tempo em diversas áreas, tais como, na detecção de fraudes financeiras, identificação de intrusão em redes industriais e corporativas e em meios de diagnósticos médicos. O objetivo desse trabalho é avaliar a aplicação de técnicas de aprendizagem de máquina para identificação de anomalias nas curvas dos dados de pressão e vazão obtidas durante o acionamento de válvulas gavetas em equipamentos submarinos de produção de óleo e gás. Numa curva de assinatura de válvula gaveta há cinco pontos importantes para abertura e quatro para o fechamento que caracterizam o comportamento da válvula e permitem identificar o desempenho e possíveis restrições operacionais. Foram construídos modelos de classificação utilizando as técnicas de SVM, PCA, Isolation Forest e Autoenconder, e como medida de desempenho foi utilizada a área da curva de ROC (AUC). Os resultados do melhor classificador foi o modelo com ensemble de três isolation forests.

ABSTRACT

Anomaly detection has been applied in several research areas, such as, financial fraud detection, intrusion identification in industrial and corporate networks and medical diagnostics. In this work, we will evaluate the application of machine learning techniques to identify anomalies in pressure and flow curves obtained during the actuation of subsea gate valves in oil and gas production equipments. In a gate valve signature curve there are five points for opening and four for closing that characterize its behavior and allow us identify performance and operational restrictions. Classification models were built using SVM, PCA, Isolation Forest and Autoenconder, and the area of the ROC curve (AUC) as a performance measure. The best classifier was the model with an ensemble of three isolation forests.

Lista de Figuras

Lista de Tabelas

Sumário

1 INTRODUÇÃO

A produção de óleo e gás no Brasil se concentra no ambiente offshore. Segundo a Agência Nacional do Petróleo, Gás Natural e Biocombustíveis – ANP (2020), em junho de 2020 a produção offshore foi responsável por 96,8% do petróleo e 85,8 % do gás natural de toda a produção brasileira. Mesmo em um cenário de recessão econômica, o volume de hidrocarbonetos produzidos foi 17,8% e 15,6% superior, respectivamente para petróleo e gás natural, na comparação com o mesmo mês de 2019.

Para viabilizar a produção segura e eficiente no ambiente offshore, há uma área de conhecimento específica, denominada engenharia submarina, cujo objetivo é o desenvolvimento dos sistemas de produção submarinos. Os sistemas de produção submarinos têm alguns aspectos únicos relacionados à inacessibilidade para instalação e manutenção, dessa forma exige equipamentos complexos e não usuais à indústria. Assim, monitorar e detectar comportamentos anormais ou anômalos é fundamental para se antecipar à eventos que possam restringir o volume de produção de óleo e gás ou reduzir o nível de segurança operacional.

Os principais equipamentos submarinos são as Árvores de Natal Molhada (ANM) e manifolds. A ANM é o equipamento instalado na cabeça do poço constituído de válvulas e demais acessórios que possibilita controlar o fluxo de fluidos entre o poço e a a unidade estacionária de produção (UEP), sejam estes produzidos ou para injeção no reservatório offshore, mantendo alta disponibilidade e segurança.

A norma API 6A (2019) define a árvore de natal como sendo o equipamento composto de adaptadores para cabeça do poço, válvulas, conexões, conectores mecânicos e hidráulicos e válvulas chokes, conectado na cabeça do suspensor da coluna de produção e utilizado para controlar a produção do poço. Essa definição, engloba qualquer equipamento instalado na cabeça do poço com objetivo de produzir óleo ou gás.

Nas ANMs também são instalados sensores, principalmente de pressão e temperatura. Essa configuração pode ser diferente de conforme premissas de projeto. Contudo, nos projetos mais modernos é possível acompanhar as pressões e vazões do fluido de atuação das válvulas e dos fluidos de processo, além de outras variáveis que monitoram as condições operacionais dos equipamentos.

Atualmente, há dezenas de poços offshore com equipamentos submarinos gerando dados continuamente. Contudo, identificou-se que há uma lacuna na

utilização destes dados para gerar conhecimento útil e agregar valor ao negócio, como por exemplo, utiliza-los para prever ou predizer anomalias durante o acionamento de válvulas submarinas do tipo gaveta, componente fundamental para garantir a extração de óleo e gás de maneira segura.

Anomalia é uma observação que se desvia tanto das outras observações a ponto de levantar suspeitas de que foi gerada por um mecanismo diferente (HAWKINS, 1980). Em alguns trabalhos o termo *outlier* também é utilizado como sinônimo, enquanto que (AGGARWAL, 2013) define anomalia como um subconjunto de *outliers*. Neste trabalho será utilizada o termo anomalia por ser uma definição mais precisa para caracterizar os dados de interesse dentro do conjunto de *outliers*.

De forma geral, uma abordagem para a maioria dos algoritmos de detecção de anomalia seguem três etapas: criar um modelo que representa o comportamento normal dos dados, calcular um valor para cada registro dos dados que mede a distância para o comportamento considerado normal e se esse valor for acima do limite especificado classifica-se esse registro como uma anomalia.

O modelo que representa os dados pode ser obtido a partir de técnicas supervisionadas, se houver disponível exemplos dos dados normais e anômalos, ou técnicas não supervisionadas ou probabilidade e estatística, se não houver informação a priori das classes normais e anômalas nos dados.

A detecção de anomalias tem aplicação há bastante tempo em diversas áreas, tais como, na detecção de fraudes financeiras, identificação de intrusão em redes industriais e corporativas e em meios de diagnósticos médicos.

Vários algoritmos tem sido propostos em trabalhos recentes com objetivo de detectar anomalias, mas estes métodos não foram projetados para lidar com dados de alta dimensionalidade (AGGARWAL; YU, 2001). Isso pode ser um problema quando se depara com dados reais que tem como característica altas dimensões que podem chegar na ordem de milhares. Neste trabalho, por exemplo, a dimensão dos dados será próxima de quatrocentos.

Com o aumento da dimensionalidade, muitos dos métodos convencionais de detecção de anomalias não funcionam de maneira muito eficaz. Em um cenário com alta dimensão, os dados tornam-se esparsos e as verdadeiras anomalias ficam mascaradas pelo efeito do ruído introduzido pelas várias dimensões irrelevantes para o problema, quando analisados com todas as dimensões (AGGARWAL, 2017).

Em (JI et al., 2020) foi proposto uma abordagem inteligente para diagnóstico de falhas em válvulas hidráulicas. Nesta abordagem, o conjunto de dados das amostras alimenta simultaneamente uma rede *Long Short Term Memory* (LSTM), uma rede neural convolucional (CNN) e uma *Random Forest* (RF). Todas as probabilidades são construídas como funções de atribuição de probabilidade básica (BPA) e são posteriormente calculadas no processo de fusão de informações.

No artigo (BOUCHET; PETROVSKI, 2014) foram avaliadas técnicas de aprendizagem de máquina, em conjunto com modelos matemáticos e estatísticos, para avaliar as condições operacionais do sistema de controle submarina a partir dos dados dos sensores. A arquitetura proposta utilizou redes neurais, *principal component analysis* (PCA) e support vector machine (SVM).

(NADEMI; VANFRETTI; PRETLOVE, 2020) desenvolveu um modelo estatístico para detecção de falhas em sistemas de distribuição elétrica submarino. A metodologia foi construída a partir de modelos ocultos de Markov em três fases, sendo que na última foram utilizados dados em tempo real para prever alguma degradação não observada nas condições atuais das cargas no sistema.

(ATIF QURESHI et al., 2019) projetou um *valve health identification* (VHI) para auxiliar a equipe de manutenção no monitoramento baseado em condição das válvulas. O VHI aborda o problema de identificação da condição de duas maneiras: supervisionada que prevê a falha iminente da válvula e não supervisionada que identifica e sinaliza as anomalias, ou seja, um comportamento incomum da válvula. Enquanto a abordagem supervisionada é adequada para válvulas com longo histórico operacional, a não supervisionada é adequada para válvulas sem histórico.

Na fase de estudo bibliográfico foi realizado um levantamento bibliométrico na base Scopus, cujo resultado é apresentado nas figuras a seguir da pesquisa. Os dados utilizados foram obtidos em 14/09/2020.

Os termos utilizados para busca de trabalho está representado na expressão:

TITLE-ABS-KEY:
((*fault* OR *anomaly*) AND
(*detection*) AND
(*valve**) AND
(*artificial* OR *intelligence* OR *machine* OR *learning* OR *algorithm*))

De forma geral, pode-se afirmar que a pesquisa bibliométrica buscou unir três conjuntos de temas amplos: detecção de anomalia ou falha, válvulas e inteligência artificial ou aprendizagem de máquina. A interseção dos três conjuntos resultou em 387 documentos entre artigos de periódicos e conferência, livros e revisão.

Os gráficos, apresentados na Figura 1.1, Figura 1.2 e Figura 1.3, mostram a quantidade de trabalhos publicados por ano, por país e citações. Observa-se, na Figura 1.1, que a partir de 2010 o volume de publicações nesses temas aumentou substancialmente, saindo de um platô de 10 publicações por ano para a faixa de 30 a 35 publicações por ano.

Na Figura 1.2 estão os números por países de origem das pesquisas. China, EUA e Reino Unido lideram o ranking e são responsáveis por quase 60% dos trabalhos publicados e o Brasil aparece em sexto, com 11 publicações. Já na Figura 1.3, o número de citações tem um comportamento similar ao de publicações por ano, com um atraso de três anos, apresentando um salto expressivo a partir de 2013 na quantidade de citações dos trabalhos.

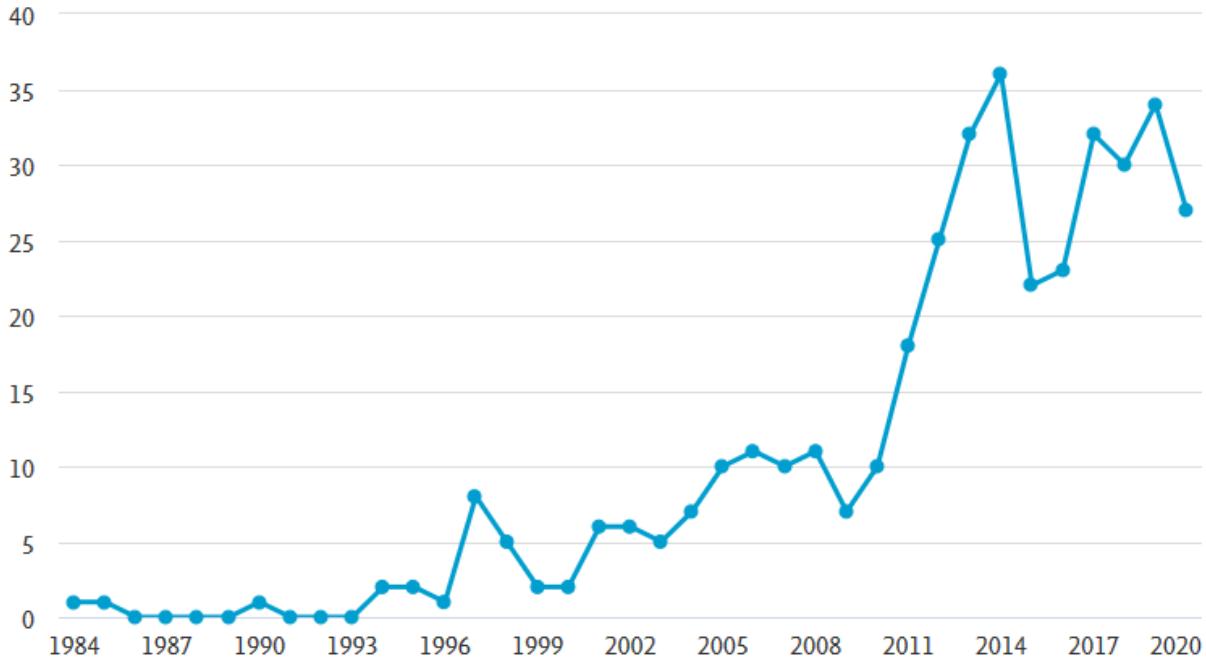


Figura 1.1 – Quantidade de trabalhos por ano.

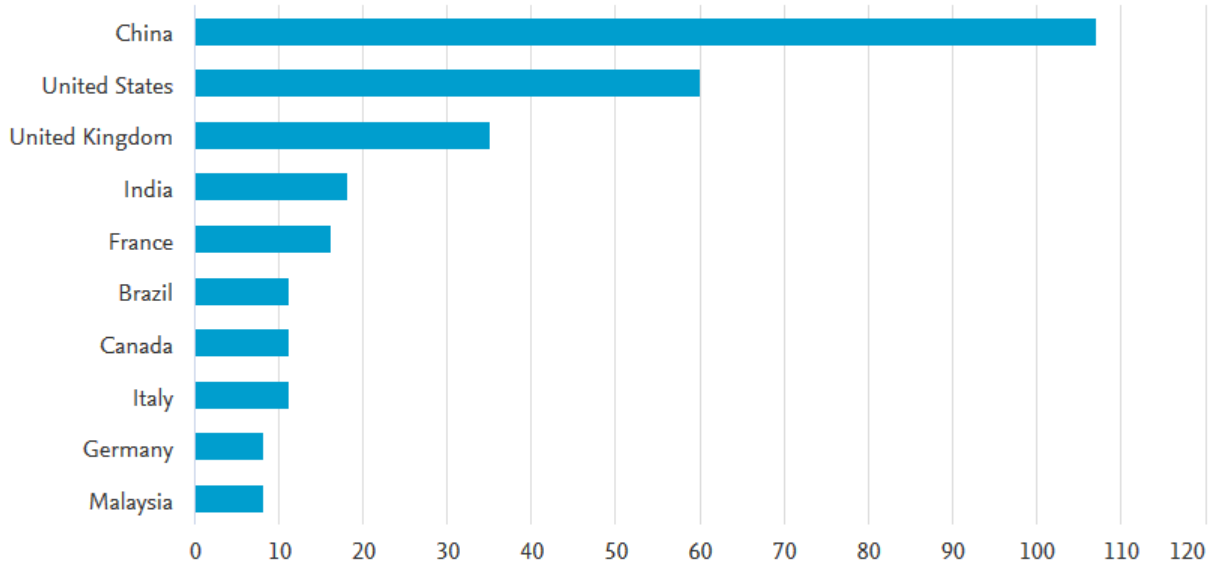


Figura 1.2 – Quantidade de trabalhos por país.

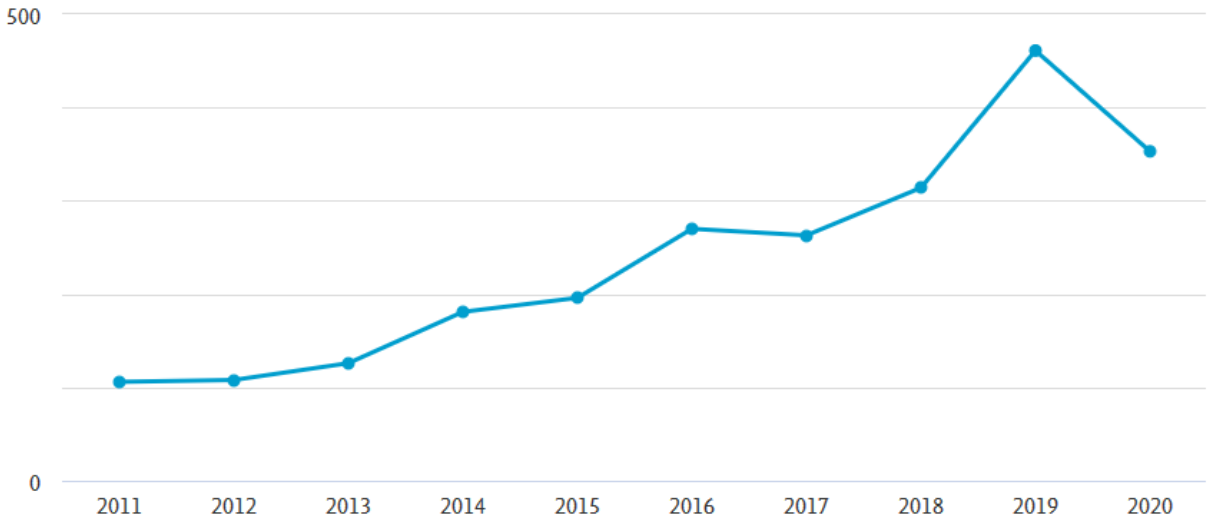


Figura 1.3 – Quantidade de citações por ano.

A Tabela 1.1 apresenta os dez trabalhos mais citados encontrados na pesquisa bibliométrica. Há trabalhos recentes e outros publicados há mais de vinte anos. Outro ponto de destaque é que nessa lista não há nenhum que trate especificamente de questões relacionadas à sistemas ou equipamentos submarinos.

A revisão bibliográfica constatou que a literatura científica recente vem estudando cada vez mais a aplicação de técnicas e métodos de aprendizagem de máquina, com destaque para as redes neurais, SVM e PCA, para identificação de anomalias. Entretanto, a interseção dessa área com engenharia submarina não há alto volume de produção científica.

Tabela 1.1 – 10 trabalhos mais citados da pesquisa bibliométrica.

Título	Autor	Ano	Fonte	Total
An approach to fault diagnosis of reciprocating compressor valves using Teager-Kaiser energy operator and deep belief networks	Tran, V.T., Althobiani, F., Ball, A.	2014	Expert Systems with Applications 41(9), pp. 4113-4122	241
A statistical, Rule-Based fault detection and diagnostic method for vapor compression air conditioners	Rossi, T.M., Braun, J.E.	1997	HVAC and R Research 3(1), pp. 19-37	156
Detection and diagnosis of oscillation in control loops	Thornhill, N.F., Hägglund, T.	1997	Control Engineering Practice 5(10), pp. 1343-1354	131
Fault detection and diagnosis for buildings and HVAC systems using combined neural networks and subtractive clustering analysis	Du, Z., Fan, B., Jin, X., Chi, J.	2014	Building and Environment 73, pp. 1-11	112
Mechanical fault diagnosis based on redundant second generation wavelet packet transform, neighborhood rough set and support vector machine	Li, N., Zhou, R., Hu, Q., Liu, X.	2012	Mechanical Systems and Signal Processing 28, pp. 608-621	95
A GMDH neural network-based approach to robust fault diagnosis: Application to the DAMADICS benchmark problem	Witczak, M., Korbicz, J., Mrugalski, M., Patton, R.J.	2006	Control Engineering Practice 14(6 SPEC. ISS.), pp. 671-683	91
Practical model and detection algorithm for valve stiction	Kano, M., Maruta, H., Kugemoto, H., Shimizu, K.	2004	IFAC Proceedings Volumes (IFAC-PapersOnline) 37(9), pp. 859-864	86
Robust on-line fault detection diagnosis for HVAC components based on nonlinear state estimation techniques	Bonvini, M., Sohn, M.D., Granderson, J., Wetter, M., Piette, M.A.	2014	Applied Energy 124, pp. 156-166	64
A new approach to fetal echocardiography: Digital casts of the fetal cardiac chambers and great vessels for detection of congenital heart disease	Gonçalves, L.F., Espinoza, J., Lee, W., (...), Mazor, M., Romero, R.	2005	Journal of Ultrasound in Medicine 24(4), pp. 415-424	64
Assumption-free anomaly detection in time series	Wei, L., Kumar, N., Lolla, V., (...), Lonardi, S., Ratanamahatana, C.A.	2005	Proceedings of the International Conference on Scientific and Statistical Database Management, SSDBM pp. 237-240	61

1.1 MOTIVAÇÃO

As válvulas submarinas são componentes críticos e cruciais para a produção segura de óleo e gás, logo monitorar o desempenho e detectar anomalias é um desafio para tornar a operação mais robusta e confiável.

Os operadores devem acionar as válvulas de forma adequada e querem mantê-las em boas condições operacionais, como também desejam ser notificados caso alguma variável indique uma anomalia. Essa anomalia pode indiciar que a válvula iniciou um mecanismo de falha ainda incipiente e que exige uma investigação mais rigorosa. Com isso a correção da falha pode ser mais fácil antes que a anomalia evolua para um problema mais sério que exija uma correção complexa, custosa ou demorada.

A análise das curvas de assinatura feitas por um especialista retira do operador a informação se está operando de forma correta. Além disso, a redução do tempo entre a geração dos dados e o laudo do especialista pode ser importante na mitigação de danos gerados por uma anomalia previamente detectadas e que evolua para uma falha crítica.

A aplicação de técnicas de aprendizagem de máquina nos dados relativos aos sistemas submarinos é uma área de trabalho escassa na literatura técnico científica, como será apresentada na revisão bibliométrica. Portanto, se justifica estudar o uso dessas ferramentas que vêm trazendo ganhos de confiabilidade, previsibilidade e compreensão dos sistemas e equipamentos em outras áreas da indústria.

1.2 OBJETIVOS DO TRABALHO

O objetivo geral desse trabalho é avaliar a aplicação de técnicas de aprendizagem de máquina para identificação de anomalias nas curvas dos dados de pressão e vazão obtidas durante o acionamento de válvulas gavetas em equipamentos submarinos de produção de óleo e gás.

Os objetivos específicos são:

1. Realizar análise exploratória dos dados históricos;
2. Aplicar técnicas de detecção de anomalias;
3. Avaliar resultados e propor aplicações práticas dos melhores resultados obtidos.

1.3 DESCRIÇÃO DO TRABALHO

O desenvolvimento dessa monografia envolveu cinco etapas: pesquisa bibliográfica e estudo sobre técnicas de detecção de anomalias; obtenção de dados históricos de assinatura de válvulas gavetas, modelagem e desenvolvimento de arquiteturas adequadas para identificar as anomalias, avaliação dos desempenhos e escolha do modelo que melhor represente os dados.

Esta monografia está dividida em cinco capítulos adicionais:

- O capítulo 2 descreve de forma ampla o problema, a configuração física dos equipamentos e a estrutura do conjunto de dados.
- O capítulo 3 descreve a abordagem metodológica e recursos utilizados no trabalho.
- O capítulo 4 descreve detalhadamente a análise exploratória dos dados e as arquiteturas para os modelos de detecção de anomalias aplicados.
- O capítulo 5 mostra os resultados obtidos.
- Por fim, o capítulo 6 descreve as conclusões do trabalho e indica oportunidades de trabalhos futuros.

2 DESCRIÇÃO DO PROBLEMA

O sistema de produção submarino objeto desse trabalho é composto por 7 poços de produção interligados a 2 manifolds que recebem a produção dos poços e estão interligados a plataforma de produção por um duto de aproximadamente 16 km. A Figura 2.1 é uma representação esquemática simplificada de como estão interligados os equipamentos do sistema de produção, onde as ANMs coletam a produção dos poços que passam pelos manifolds que direciona o escoamento da produção para os dutos rígidos até a UEP.

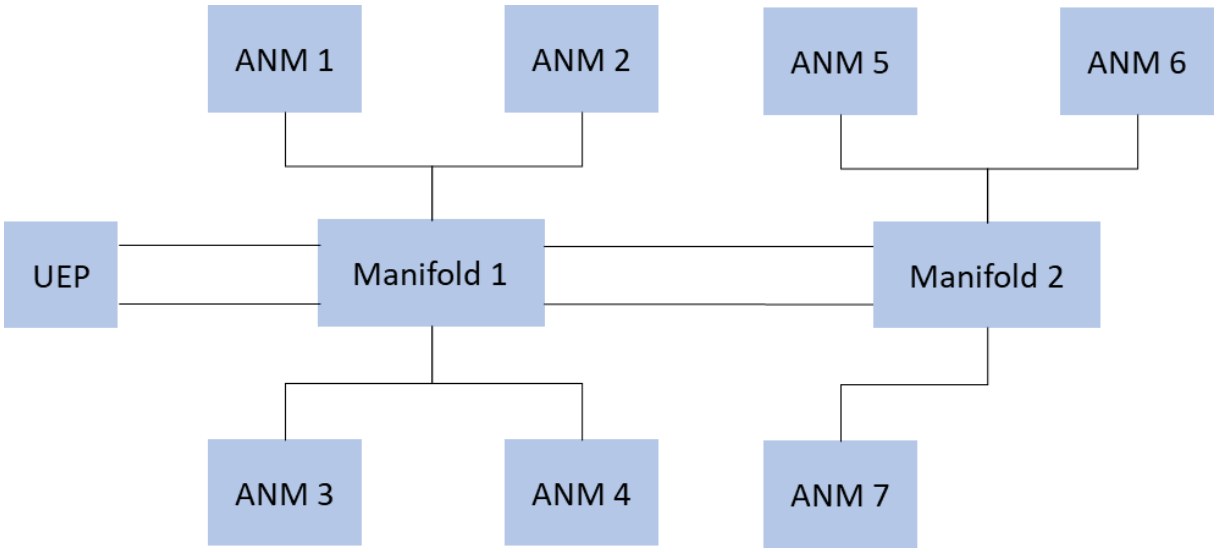


Figura 2.1 – Esquemático da configuração dos equipamentos.

Cada equipamento representando na Figura 2.1 possui válvulas do tipo gaveta que é o componente mais importante e crítico para a produção e segurança. O sistema de controle possui sensores de pressão e vazão que monitoram essas variáveis durante os comandos de abertura ou fechamento enviados pela equipe de operação a bordo da UEP.

Essas curvas de assinatura de válvula são apresentadas no sistema supervisório e, de acordo, com os limites pré-estabelecidos, geram alarmes de movimentação de válvulas sinalizando aos operadores a necessidade de observar com mais atenção determinada ANM, válvula ou comando enviado.

As válvulas mais utilizadas em sistemas submarinos, do ponto de vista construtivo, são as do tipo gaveta, esfera e choke. As duas primeiras são do tipo aberto-fechado (*on-off*) e a terceira é do tipo dosadora ou de controle. Essas válvulas

são acionadas através de energia hidráulica em seus atuadores para realizar o movimento desejado.

Uma válvula gaveta hidráulica pode ser dividida em: válvula, bonnet e atuador hidráulico. A válvula propriamente dita é o componente que possui a gaveta, ou seja, é a parte que interrompe ou permite o fluxo do fluido de produção, o bonnet é responsável por realizar a ligação entre a válvula e o atuador hidráulico e o atuador hidráulico é o responsável por realizar a abertura e o fechamento da válvula (MASHIBA, 2011).

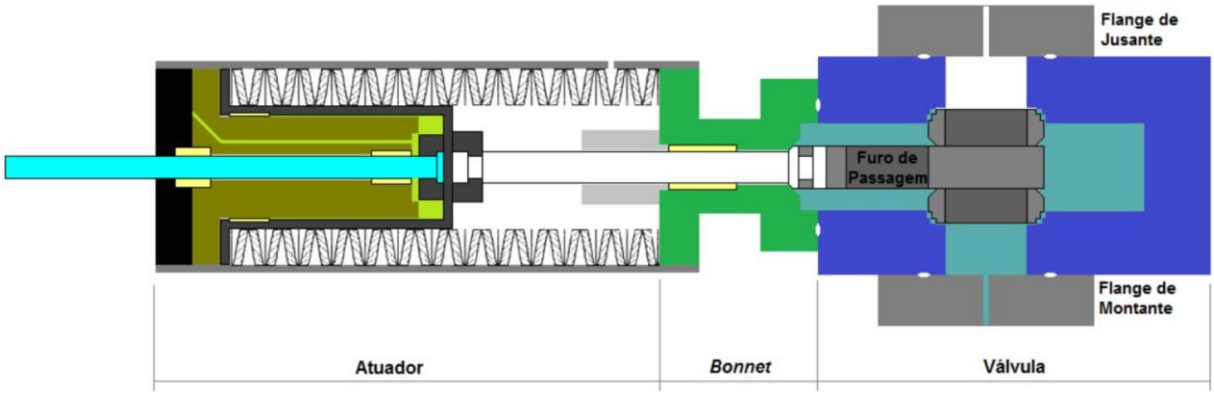


Figura 2.2 - Válvula gaveta com atuador hidráulico com retorno por mola.

Fonte: (MASHIBA, 2011).

Numa curva de assinatura de válvula gaveta há cinco pontos importantes para abertura e quatro para o fechamento que caracterizam o comportamento da válvula e permitem identificar o desempenho e possíveis restrições operacionais. São eles, para abertura: início do movimento da haste de override do atuador, início do movimento da gaveta, início de comunicação, completa equalização e final de avanço do atuador; e para o fechamento: início do retorno do atuador, fim de comunicação, completo diferencial e fim de retorno do atuador (EUTHYMIOU, 2001).

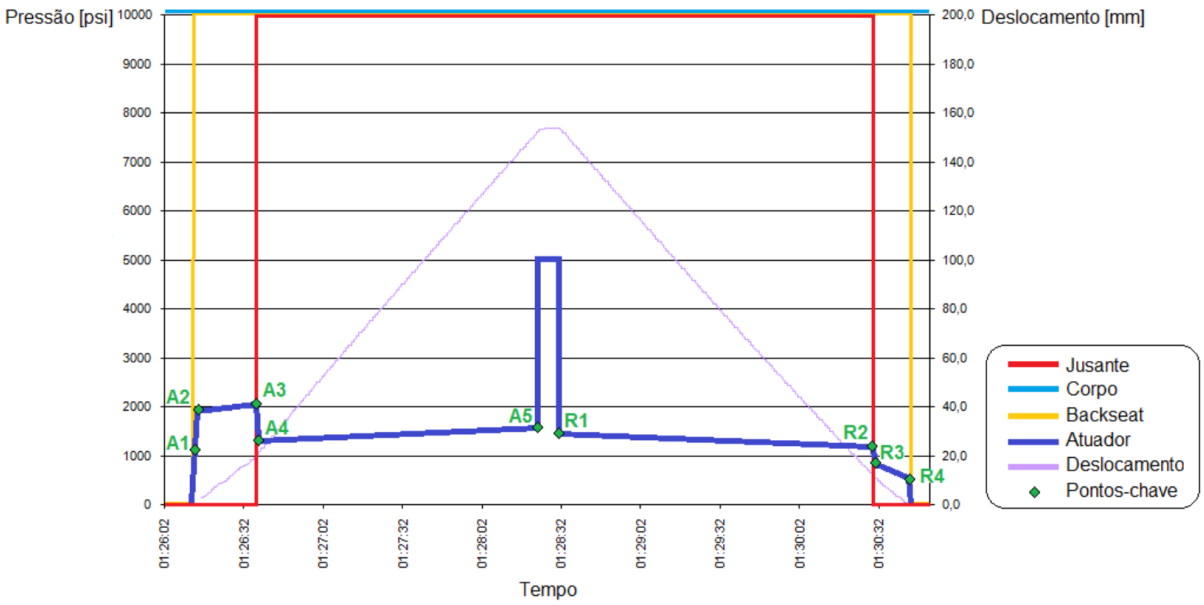


Figura 2.3 – Curva de atuação característica (assinatura) de uma válvula gaveta.

Fonte: (EUTHYMIOU, 2001).

A curva apresentada na Figura 2.3 é uma assinatura de válvula gaveta realizada em laboratório. Nela estão destacados os cinco pontos-chaves para abertura (A1 a A5) e os quatro para fechamento (R1 a R4). Essa curva é denominada assinatura devido ao fato que independente de ser ou não do mesmo fabricante, cada válvula possui sua curva característica (MASHIBA, 2011).

2.1 DESCRIÇÃO DOS DADOS

Todos os experimentos realizados neste trabalho utilizaram o conjunto de dados de assinatura de válvulas gavetas submarinas que foi disponibilizado pela Petróleo Brasileiro S.A. – Petrobras. Esse conjunto de dados não é público.

Os dados são coletados nas UEPs a cada quinze dias e passam por um especialista que emite um laudo sobre cada registro de atuação de válvula, só depois dessa fase esses registros são recebidos pelas áreas técnicas e operacionais. Esse processo acaba gerando vários arquivos em formato de planilha eletrônica que foi consolidado por Medeiros (2019) em um único arquivo tipo csv, portanto detalhes sobre a extração dos dados podem ser consultados nesta referência.

Os dados consolidados foram organizados em arquivo contendo 131 colunas representando as características. As dez primeiras contém dados que identificam a válvula, poço, data, tipo de comando enviado, etc. a última coluna é o laudo emitido

pelo especialista e as colunas da posição onze até a cento e trinta contém os 120 pontos de amostra da variável obtido durante o acionamento da respectiva válvula.

O histórico de acionamentos obtido possui 19164 registros. Como são monitoradas duas variáveis de pressão e uma de vazão, cada atuação de válvula é representado por uma tríade de registro nos dados, por exemplo, os registros da posição 0, 1 e 2 são relativos a uma única atuação de válvula. Assim, há 6388 comandos de acionamentos que serão utilizados como dados de entrada.

A seguir uma descrição das colunas categóricas:

1- Qual poço a válvula está relacionada, caso seja uma válvula de ANM, e qual manifold a válvula está relacionada caso seja uma válvula de manifold (textual);
2- Nome do poço que a ANM está vinculada, ou nome do conjunto manifold que o manifold está vinculado (textual);
3- Identificador da válvula para a empresa (textual);
4- Nome comum da válvula (textual);
5- Identificador da válvula para o sistema (numérico);
6- Código da operação (numérico);
7- Tipo de operação: comando para abrir ou fechar a válvula (textual);
8- Posição da válvula: aberta ou fechada (binário);
9- Timestamp: data e hora de envio do comando (timestamp);
10- Tempo de amostragem das variáveis temporais (numérico); e
11- Laudo do técnico especialista (textual)

2.1.1 Análise Exploratória dos Dados

Os dados coletados são referentes a 80 válvulas, onde foram realizados 6388 comandos de atuação de abertura ou fechamento, sendo 3446 e 2942 comandos respectivamente. Do total de válvulas, 56 pertencem as ANMs instaladas nos poços e 24 são dos dois manifolds de coleta de gás.

A coluna alvo (objetivo) é o laudo do técnico especialista, que por ser um campo textual livre, foi necessário realizar um pré-processamento. Nos dados originais há, nessa amostra, 187 categorias distintas na coluna laudo, porém na prática os acionamentos de válvulas podem ser classificados em 3 categorias genéricas: acionamento sem problema (ok), comando enviado para válvula já na posição

(already_on_position) e acionamento com algum desvio do comportamento esperado (alert). Além disso, foi criada a categoria nulo para representar os registros cujo campo laudo não foi possível enquadrar em nenhum dos três anteriores.

A Figura 2.4 é a distribuição dos dados pelas categorias reformuladas do campo laudo (ok = 63,46%, already_on_position = 27,28%, alert = 6,90% e nulo = 2,36%). Para fins de identificação de anomalias, a categoria ok e already_on_position podem ser consideradas como a categoria com comportamento normal.

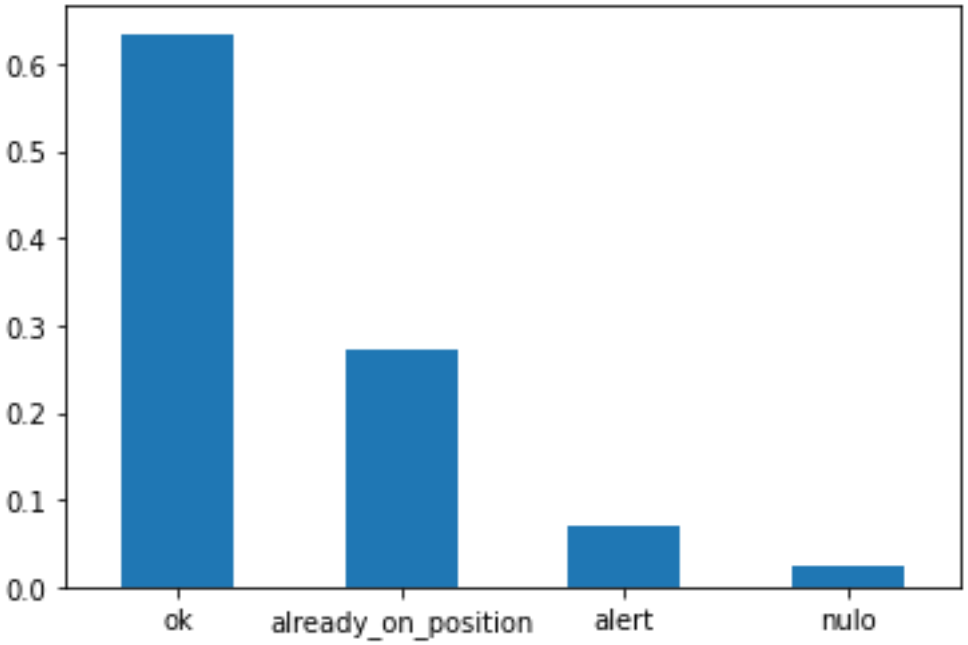


Figura 2.4 – Distribuição dos acionamentos de válvula pelo laudo.

Com relação a distribuição dos dados no tempo, o primeiro registro foi armazenado em 14/09/2009 e o último em 08/08/2018. Os gráficos das Figura 2.5 e Figura 2.6 mostra a distribuição da quantidade dos acionamentos por dia da semana e por mês do ano em números absolutos. Os valores percentuais dos acionamentos por dia da semana e mês do ano estão na Figura 2.7 e na Figura 2.8.

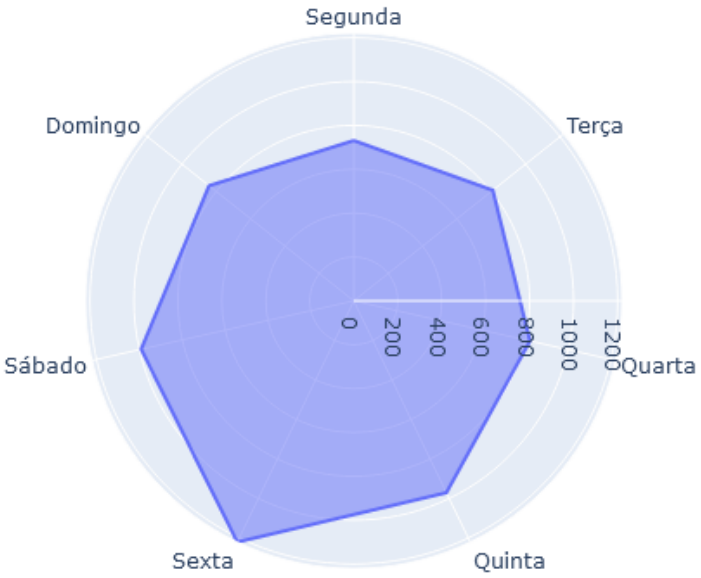


Figura 2.5 – Distribuição dos acionamentos de válvula por dia da semana.

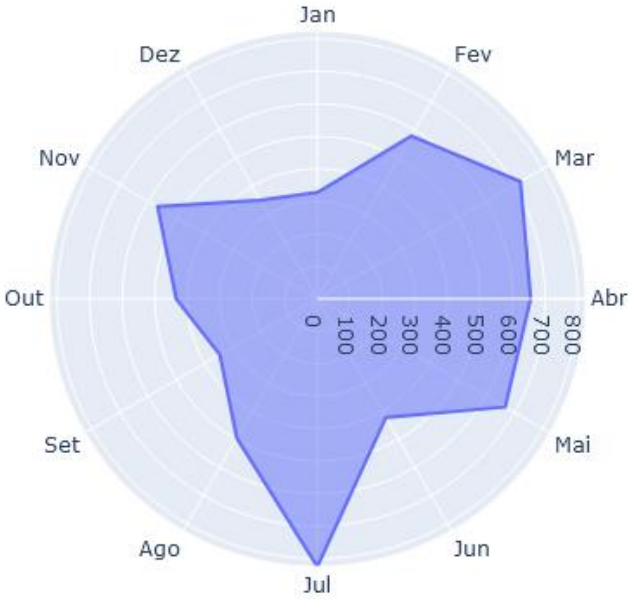


Figura 2.6 – Distribuição dos acionamentos de válvula por mês do ano.

Sexta	19.113964
Sábado	15.529117
Quinta	15.200376
Domingo	13.180964
Quarta	12.867877
Terça	12.664371
Segunda	11.443331

Figura 2.7 – Percentual dos acionamentos de válvula por dia da semana.

Jul	12.914840
Mar	11.286788
Mai	10.441453
Abr	10.269255
Fev	9.048215
Nov	8.876018
Ago	7.748904
Out	6.793989
Jun	6.606137
Dez	5.479023
Set	5.416406
Jan	5.118973

Figura 2.8 – Percentual dos acionamentos de válvula por mês do ano.

O campo contendo o nome comum da válvula (TAG) também passou por um pré-processamento para encurta-lo, transformando-o em uma sigla. Um exemplo da transformação é a válvula com TAG `Production Master Valve 2 [M2]` que passou a ser representada apenas por M2 (sigla entre colchete). A sigla utilizada para representar a válvula, nos casos das ANMs, seguem a mesma nomenclatura da norma API 6A (2019). Essa mesma abordagem foi utilizada para encurtar os demais nomes de válvulas.

Após esse tratamento foi construído o gráfico da Figura 2.9 que mostra o número absoluto de acionamentos pelo nome da válvula. As válvulas M1 e W1 são as mais operadas, provavelmente por estarem ligadas diretamente a produção de óleo e gás e a segurança (garantia de total fechamento) do poço.

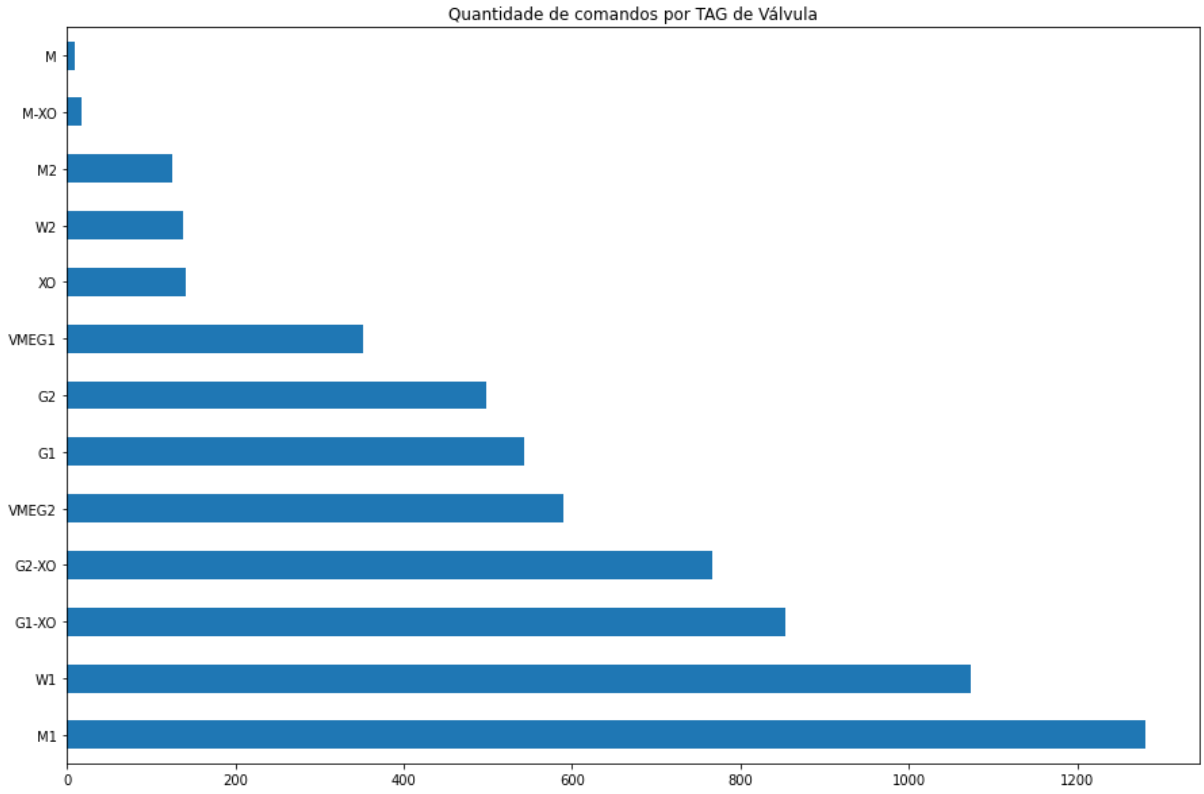


Figura 2.9 – Número de acionamentos de válvula por nome (TAG).

A Figura 2.10 traz 4 amostras com laudo OK. As quatro séries apresentam fluxo de fluido, o que permite afirmar que houve movimentação da válvula devido o preenchimento ou esvaziamento da câmara de atuação. Como a Série 2 tem o valor máximo em 0,16, enquanto que as demais ficam acima de 1,2 pode-se afirmar que a Série 2 representa o volume do atuador de uma válvula menor (2 polegadas) e as demais séries representam a válvula maior (5 polegadas). Já nos gráficos de pressão, observa-se dois perfis distintos: séries 2 e 4 e séries 1 e 3. O primeiro perfil tem comportamento que caracteriza movimentos de abertura de válvula, onde as linhas da função tem aumento rápido, uma oscilação durante o deslocamento da haste de atuação, e por fim uma estabilização próximo ao valor da pressão de operação da unidade hidráulica da plataforma (entre 3200 e 3500 psi). O segundo perfil de pressão caracteriza o movimento oposto, de fechamento, com queda abrupta da pressão da linha de função que se estabiliza em um valor próximo da pressão da linha de retorno.

Na Figura 2.11 pode-se fazer análise similar ao da Figura 2.10. Neste caso, o objetivo é mostrar algumas diferenças que podem acontecer nos perfis de pressão e fluxo quando se compara válvulas de tamanho distintos. Isso pode influenciar a

estratégia de normalização e escalonamento dos dados, na fase de preparação, que pode influenciar no resultado de determinado algoritmo.

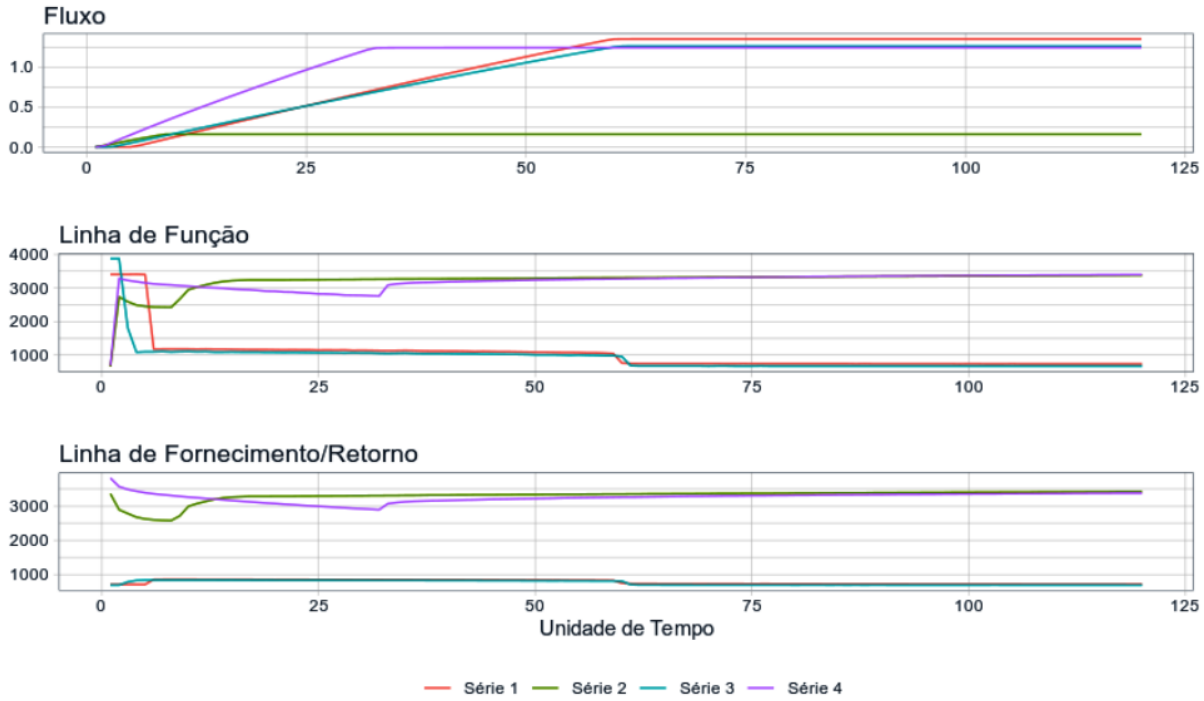


Figura 2.10 – Curvas com laudo “OK”.

Fonte: (MEDEIROS, 2019).

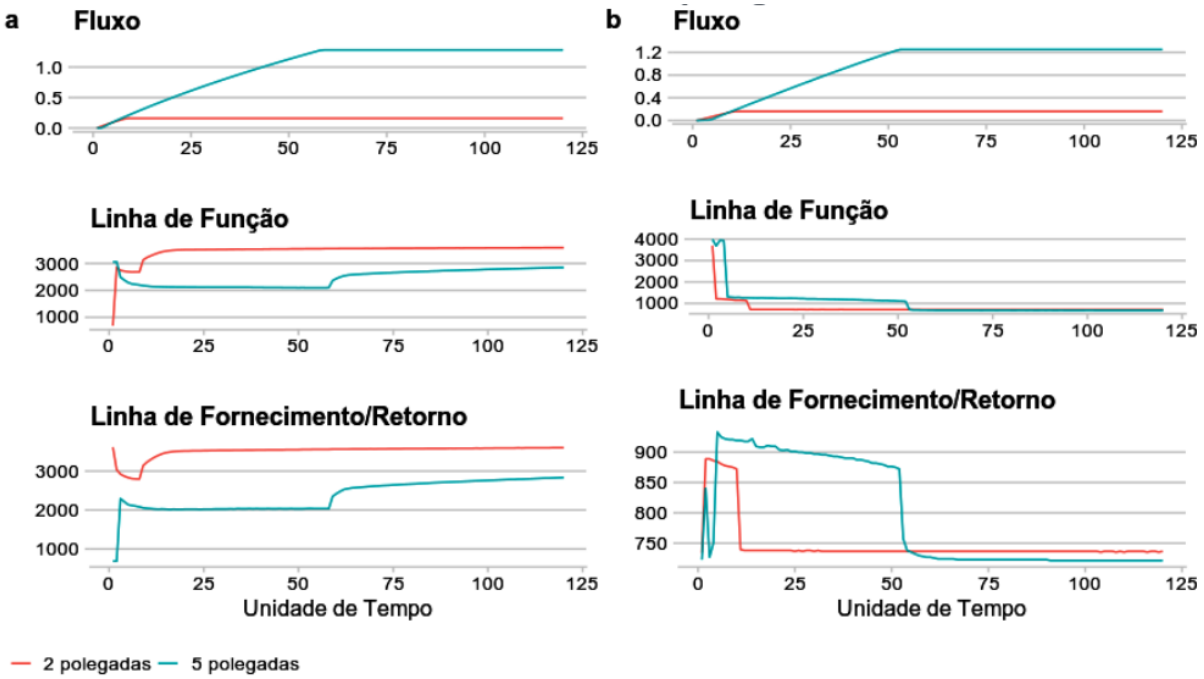


Figura 2.11 – Comparação entre operações de abertura (a) e fechamento (b) de válvulas de 5 e 2 polegadas para laudo “OK”.

Fonte: (MEDEIROS, 2019).

3 METODOLOGIA

Neste trabalho serão utilizadas técnicas de aprendizagem de máquina na detecção de anomalias. Foram escolhidos métodos que fossem mais adequados para trabalhar com dados multidimensionais (360 variáveis de entrada, nesse trabalho) e classes desbalanceadas (93% de ocorrências normais). A seguir são apresentadas os recursos utilizados e as definições das técnicas de aprendizagem de máquina, bem como da métrica para comparação do desempenho no conjunto de teste.

3.1 HARDWARE E SOFTWARE

Todas as fases práticas deste trabalho foram desenvolvidas em linguagem Python, no Google Colab, utilizando as diversas bibliotecas construídas pela comunidade para aplicações de aprendizagem de máquina e inteligência artificial, sendo os principais: Numpy, Pandas, Matplotlib, Scikit Learn e PyOD. Todos os pacotes utilizados podem ser vistos no código desenvolvido que segue apêndice.

3.2 TÉCNICAS DE CLASSIFICAÇÃO e MÉTRICA DE DESEMPENHO

3.2.1 Support Vector Machine – SVM

O método support vector machine implementa a seguinte ideia: mapeia os vetores de entrada em algum espaço de características de alta dimensão por meio de um mapeamento não linear escolhido a priori. Neste espaço, uma superfície de decisão linear é construída com propriedades que garantem a capacidade de generalização (CORTES; VAPNIK, 1995).

O SVM combina três ideias: a técnica de solução parra hiperplanos ótimos de separação das classe (que permite uma expansão do vetor de solução em vetores de suporte), a convolução do produto escalar (que estende as superfícies de solução de linear para não linear), e a noção de margens suaves (para permitir erros no conjunto de treinamento). Os detalhes matemáticos desse método estão disponíveis em (CORTES; VAPNIK, 1995).

3.2.2 Principal Component Analysis – PCA

A análise de componente principal (PCA) reduz a dimensão dos dados sem muita perda das informações. Isso é feito porque em dados com muitas dimensões, apenas alguns parâmetros dos dados contém a informação valiosa para detecção da anomalia. Essa técnica também será utilizada como forma de preparação dos dados

de entrada de outros algoritmos. Isso traz uma vantagem adicional de reduzir o tempo de execução, o que pode facilitar a aplicação em tempo real. Além disso, o PCA é utilizado na fase de exploração, porque permite criar gráficos e visualizações em duas dimensões.

Os componentes principais são combinações lineares das n variáveis aleatórias X1, X2, ..., Xn, e possuem três propriedades fundamentais: os componentes principais não estão correlacionados, o primeiro componente principal tem a maior variância, o segundo principal o componente tem a segunda maior variância e assim por diante, e a variação total em todos os componentes principais combinados é igual à variação total nas variáveis originais X1, X2, ..., Xn (SHYU et al., [s.d.]).

3.2.3 Isolation Forest

Isolation Forest é um método de detecção de anomalias que constrói um conjunto de isolation trees. As anomalias são aquelas amostras que tem um curto percurso médio no conjunto de isolation trees. Neste método há apenas duas variáveis: o número de árvores a serem construídas e o tamanho da subamostra. O desempenho de detecção converge de forma bastante rápida com um número muito pequeno de árvores e exige um pequeno conjunto de subamostragem (LIU; TING; ZHOU, 2008).

Por causa das características das amostras categorizadas como anomalias (menor proporção no conjunto de dados e característica muito diferente das amostras normais), as anomalias são isoladas mais perto da raiz da árvore, enquanto os pontos normais são isolados em ramos mais profundos da árvore. Esta característica de isolamento da árvore forma a base desse método. Mais detalhes sobre o método podem ser encontrados em (LIU; TING; ZHOU, 2008).

3.2.4 Autoencoder

Um autoencoder é um algoritmo de extração de características não supervisionado baseado em rede neural. Nesse tipo de rede, o treinamento é feito com objetivo de aprender as melhores características que são necessárias para reconstruir a entrada na sua saída, o mais próximo possível de sua entrada.

Embora reconstruir os dados possa parecer uma questão trivial, simplesmente copiando os dados para a frente de uma camada para outra, isso não é possível quando o número de neurônios da camada do central é menor que a quantidade de variáveis de entrada. Ou seja, o número de neurônicos da camada intermediária é

menor do que da camada de entrada (ou de saída). Isso resulta que os neurônios dessa camada mantêm uma representação dos dados. Portanto, este método de reconstrução tem perda (AGGARWAL, 2018).

3.2.5 Receiver Operating Characteristic – ROC

O ROC é um método quantitativo para comparar uma variável booleana de referência com a probabilidade de pertencer a uma determinada classe ou possuir certa característica. Quanto maior for esse valor, mais provável é a presença da característica em análise (PONTIUS; PARMENTIER, 2014).

A comunidade de aprendizagem de máquina usa com maior frequência a métrica denominada AUC (Area Under the Curve). Essa medida é a área da curva ROC e varia de 0 a 1 (sem unidade), onde valores maiores indicam melhores resultados da classificação.

Essa métrica é bastante questionada porque não consegue expressar toda a informação presente na curva de ROC. Mesmo sendo uma sumarização da curva ROC acaba sendo mais utilizada do que a própria curva, como foi constatado em (PONTIUS; PARMENTIER, 2014). Entretanto, a coerência do uso da AUC como métrica de desempenho do classificador é justificada em diversos estudos. Portanto, neste trabalho utilizaremos a medida da AUC, que é um resumo da curva ROC.

4 ARQUITETURA DO SISTEMA PROPOSTO

Foram simuladas algumas configurações de arquiteturas de modelos de classificação binária de anomalias. Os algoritmos utilizados são: PCA, SVM, Isolation Forest e Autoencoder. O PCA também foi aplicado na fase de análise exploratória para montar visualizações em duas dimensões, além disso, foi aplicado na redução de dimensionalidade das variáveis de entrada.

Para os dados de entrada adotou-se duas estratégias de pós-tratamento. A primeira foi normaliza-los utilizando a técnica de subtrair a média e dividir os dados pelo desvio padrão. A segunda abordagem foi realizada uma redução de dimensionalidade, utilizado o PCA, para cinco dimensões. Esse número de dimensões (cinco) foi escolhido porque preserva 99% da variância dos dados. A tabela abaixo resume os modelos construídos que tiveram os melhores resultados.

Tabela 4.1 – Modelos de classificadores de anomalias.

ID	Algoritmo	Modelo
A	SVM	OCSVM(cache_size=200, coef0=0.0, contamination=0.172, degree=3, gamma='auto', kernel='rbf', max_iter=-1, nu=0.5, shrinking=True, tol=0.001, verbose=False)
B	SVM	OCSVM(cache_size=200, coef0=0.0, contamination=0.172, degree=3, gamma='auto', kernel='rbf', max_iter=-1, nu=0.5, shrinking=True, tol=0.001, verbose=False)
C	Isolation Forest	IForest(behaviour='new', bootstrap=False, contamination=0.172, max_features=1.0, max_samples='auto', n_estimators=100, n_jobs=1, random_state=None, verbose=0)
D	Isolation Forest	IForest(behaviour='new', bootstrap=False, contamination=0.172, max_features=1.0, max_samples='auto', n_estimators=100, n_jobs=1, random_state=None, verbose=0)
E	Esemble de três Isolation Forests	IForest(behaviour='new', bootstrap=False, contamination=0.172, max_features=1.0, max_samples=200, n_estimators=100, n_jobs=1, random_state=None, verbose=0) IForest(behaviour='new', bootstrap=False, contamination=0.172,

		max_features=1.0, max_samples=100, n_estimators=100, n_jobs=1, random_state=None, verbose=0) IForest(behaviour='new', bootstrap=False, contamination=0.172, max_features=1.0, max_samples=1000, n_estimators=100, n_jobs=1, random_state=None, verbose=0)
F	Autoencoder	AutoEncoder(batch_size=32, contamination=0.172, dropout_rate=0.2, epochs=100,hidden_activation='relu', hidden_neurons=[64, 32, 32, 64], l2_regularizer=0.1, loss=<function mean_squared_error at 0x7f09317fa268>, optimizer='adam', output_activation='sigmoid', preprocessing=True, random_state=None, validation_size=0.1, verbose=1)
G	Autoencoder	AutoEncoder(batch_size=32, contamination=0.172, dropout_rate=0.2, epochs=100, hidden_activation='relu', hidden_neurons=[5, 2, 2, 5], l2_regularizer=0.1, loss=<function mean_squared_error at 0x7f09317fa268>, optimizer='adam', output_activation='sigmoid', preprocessing=True, random_state=None, validation_size=0.1, verbose=1)
H	PCA	PCA(contamination=0.172, copy=True, iterated_power='auto', n_components=5, n_selected_components=None, random_state=None, standardization=True, svd_solver='auto', tol=0.0, weighted=True, whiten=False)

5 RESULTADOS

Os dados de entrada foram preparados de três maneiras. A primeira foi através da centralização e o dimensionamento independentemente em variável de entrada fazendo a subtração da média e dividindo pelo desvio padrão. A segunda foi utilizando a redução de dimensionalidade (PCA) de 360 para 5. Esse número de dimensões foi escolhido porque consegue manter 99% da variância dos dados. A terceira foi a redução de dimensionalidade após o escalonamento dos dados. Essa terceira não será apresentada porque resultou em resultados piores que a segunda forma.

Os principais resultados obtidos são apresentados em resumo nas Tabela 5.1 e Tabela 5.2. A primeira tabela é um resumo do modelo, a maneira como os dados de entrada fora pós-processado, o algoritmo (método) de classificação e a AUC no conjunto de teste. Assim, o melhor modelo, a partir desse critério, foi o Modelo E com AUC igual a 0.7148. Esse resultado foi obtido utilizando os dados de entrada escalonados, sem redução de dimensionalidade, e um ensemble de três classificadores que utilizam o método Isolation Forest.

Tabela 5.1 – Resumo dos resultados no conjunto de teste.

ID	Dados de Entrada	Método Classificação	AUC no Teste
A	Escalonado*	SVM	0.4895
B	Redução para 5D com PCA	SVM	0.6476
C	Escalonado*	Isolation Forest	0.6846
D	Redução para 5D com PCA	Isolation Forest	0.6331
E	Escalonado*	Ensemble com 3 Isolation Forests	0.7148
F	Escalonado*	Autoencoder	0.4832
G	Redução para 5D com PCA	Autoencoder	0.5015
H	Escalonado*	PCA	0.4826

*dados escalonados subtraindo a média e dividindo pelo desvio padrão

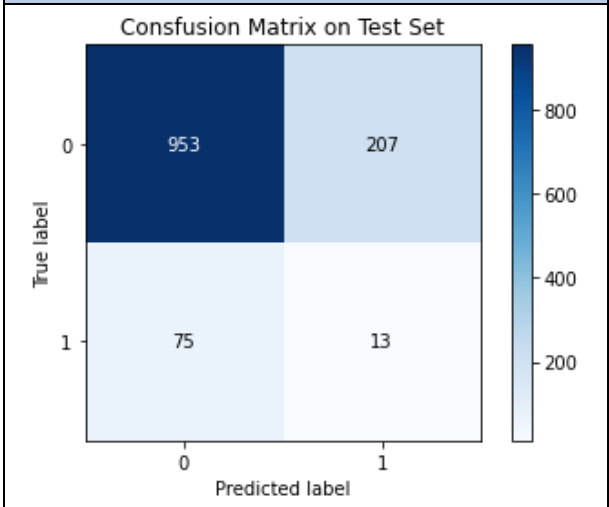
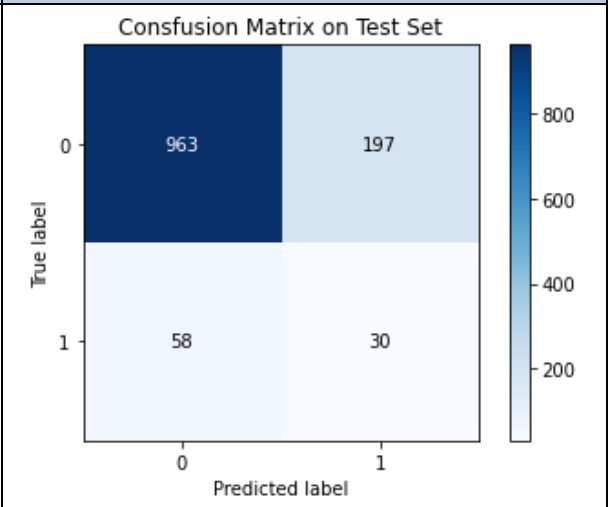
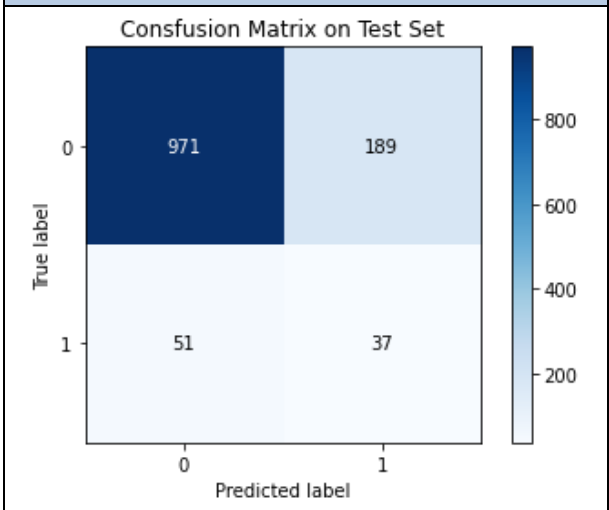
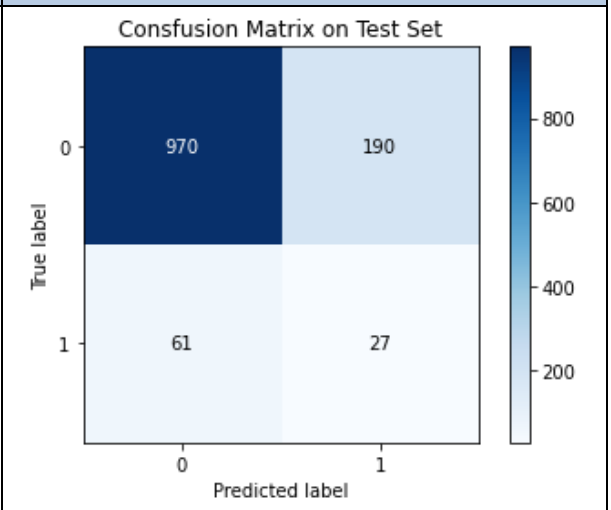
Os modelos construídos com SVM apresentam melhores resultados quando se utiliza dados após a redução de dimensionalidade. Já os modelos com autoenconder não se mostraram bons classificadores das anomalias, assim como o modelo que utiliza apenas PCA.

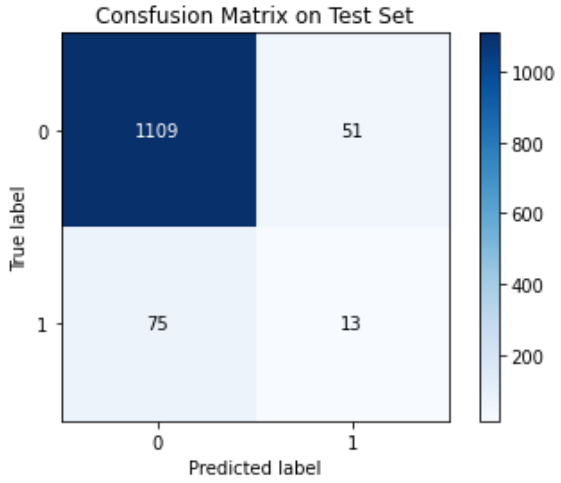
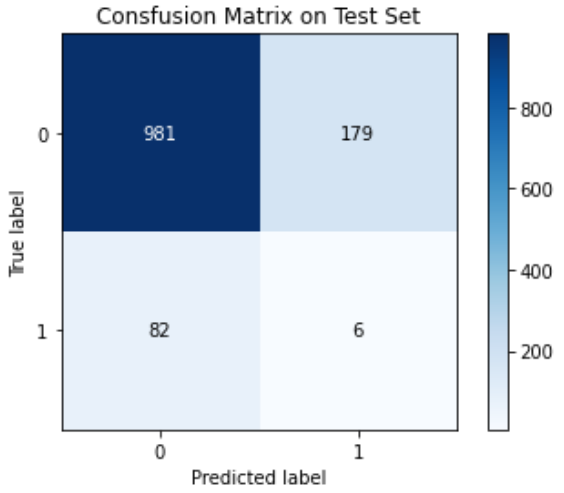
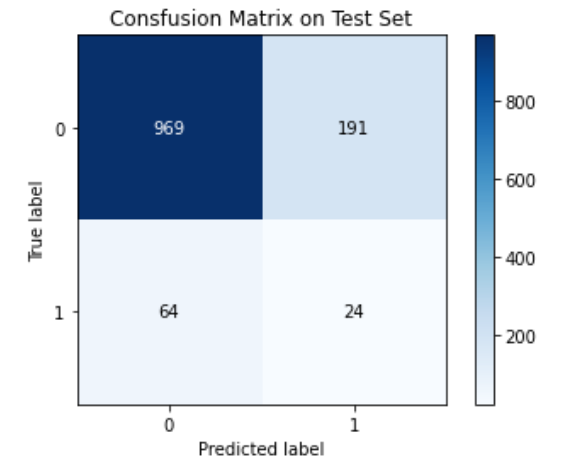
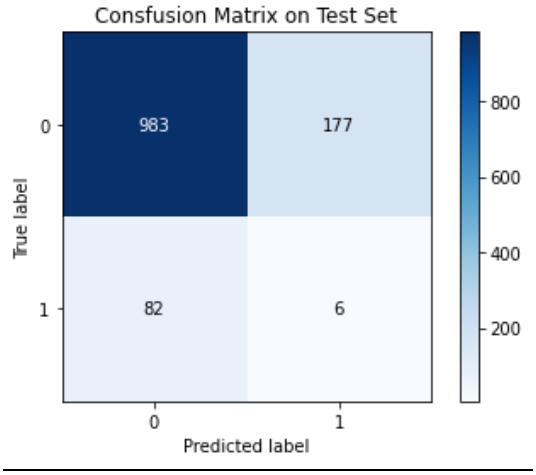
A Tabela 5.2 contém as matrizes de confusão dos oito modelos apresentados no conjunto de teste. O Modelo C foi o que identificou a maior quantidade anomalias (maior número de verdadeiro negativo) também utilizou a isolation forest como técnica de classificação. Nesse modelo, a métrica de AUC foi a segunda melhor.

Os Modelos F e H foram os que identificaram o menor número de anomalias. Já com relação ao número de falsos negativos, todos os modelos, exceto o Modelo E, diagnosticaram mais de 177 ocorrências. Esse número é bastante alto se considerar que há 1160 registros de condição normal no conjunto de teste.

Em um sistema de supervisão e automação, uma grande quantidade de falso negativos pode levar a equipe de operação a desacreditar os alarmes/sinalizações de anomalias, pois na maioria das vezes serão alarmes falsos. Nesse sentido, analisando a matriz de confusão constata-se que o melhor modelo é o Modelo E.

Tabela 5.2 – Matrizes de confusão dos modelos.

Modelo A	**Modelo B**
	
Modelo C	**Modelo D**
	

Modelo E	Modelo F
	
Modelo G	**Modelo H**
	

Os modelos propostos conseguiram identificar as anomalias, entretanto, para que seja implementando numa situação real carece de melhorias para se seja possível separar melhor as anomalias. Além disso, todos os modelos apresentados foram bastante rápidos na execução do algoritmo de classificação, assim não se espera problemas de desempenho no hardware atual do sistema, caso algum deles seja posto em produção.

6 CONCLUSÕES E TRABALHOS FUTUROS

Neste trabalho foram utilizados dados coletados durante o acionamento de válvulas submarinas para classifica-los em duas categorias: normal e anomalia. Foi realizada uma extensiva análise exploratória, seguida de uma etapa de tratamento e preparação dos dados para identificar inconsistências e tornar o conjunto de dados adequados para utilização dos métodos de classificação.

Foram avaliadas a aplicação das técnicas SVM, PCA, Isolation Forest e Autoenconder para a classificação de anomalias nas curvas de pressão e vazão obtidas durante o acionamento de válvulas gavetas submarinas. Diversos modelos foram construídos e oito configurações foram apresentadas, sendo que o melhor desempenho, obtido pelo maior valor de AUC, foi o modelo formado por um ensemble de três isolation forests. O modelo que identificou a maior quantidade anomalias (maior número de verdadeiro negativo) também utilizou a isolation forest como técnica de classificação.

No entanto, os valores obtidos para AUC de todos os modelos ainda não podem ser considerados bons, quando se observa as matrizes de confusão, pois ainda há muitas anomalias que não são identificadas corretamente. Portanto, para inserir um modelo automático de classificação online no sistema de supervisão da planta ainda se faz necessário estudos adicionais para melhorar esse desempenho.

Referências Bibliográficas

AGGARWAL, C. C. **Outlier Analysis**. New York: Springer-Verlag, 2013.

AGGARWAL, C. C. **Outlier Analysis**. Cham: Springer International Publishing, 2017.

AGGARWAL, C. C. **Neural Networks and Deep Learning: A Textbook**. Cham: Springer International Publishing, 2018.

AGGARWAL, C. C.; YU, P. S. Outlier Detection for High Dimensional Data. p. 10, 2001.

ANP. **Boletim da Produção de Petróleo e Gás Natural**, jun. 2020. Disponível em: <http://www.anp.gov.br/arquivos/publicacoes/boletins-anp/producao/2020-06-boletim.pdf>. Acesso em: 1 set. 2020

API 6A. **Specification for Wellhead and Tree Equipment**. Disponível em: <https://www.api.org:443/products-and-services/standards/important-standards-announcements/spec-6a>. Acesso em: 1 set. 2020.

ATIF QURESHI, M. et al. VHI: Valve Health Identification for the Maintenance of Subsea Industrial Equipment. In: BREFELD, U. et al. (Eds.). . **Machine Learning and Knowledge Discovery in Databases**. Lecture Notes in Computer Science. Cham: Springer International Publishing, 2019. v. 11053p. 668–671.

BOUCHET, F.; PETROVSKI, A. **Adaptive fault detection tool for real-time integrity monitoring of Subsea Control Systems**. 2014 IEEE International Symposium on Innovations in Intelligent Systems and Applications (INISTA) Proceedings. **Anais**... In: 2014 IEEE INTERNATIONAL SYMPOSIUM ON INNOVATIONS IN INTELLIGENT SYSTEMS AND APPLICATIONS (INISTA). Alberobello, Italy: IEEE, jun. 2014Disponível em: <http://ieeexplore.ieee.org/document/6873592/>. Acesso em: 24 jul. 2020

CORTES, C.; VAPNIK, V. Support-vector networks. **Machine Learning**, v. 20, n. 3, p. 273–297, set. 1995.

EUTHYMIOU, E. J. **Metodologia para testes funcionais em válvulas submarinas**. Dissertação—Rio de Janeiro, Brasil: Universidade Federal do Rio de Janeiro, 2001.

HAWKINS, D. **Identification of Outliers**. [s.l.] Springer Netherlands, 1980.

JI, X. et al. An intelligent fault diagnosis approach based on Dempster-Shafer theory for hydraulic valves. **Measurement**, p. 108129, jun. 2020.

LIU, F. T.; TING, K. M.; ZHOU, Z.-H. **Isolation Forest**. 2008 Eighth IEEE International Conference on Data Mining. **Anais**... In: 2008 EIGHTH IEEE INTERNATIONAL CONFERENCE ON DATA MINING (ICDM). Pisa, Italy: IEEE, dez. 2008Disponível em: <http://ieeexplore.ieee.org/document/4781136/>. Acesso em: 1 nov. 2020

MASHIBA, M. H. DA S. **A influência dos parâmetros de operação e projeto no desempenho de atuação hidráulica de válvulas submarinas do tipo gaveta**. Dissertação—Rio de Janeiro, Brasil: Universidade Federal do Rio de Janeiro, 2011.

MEDEIROS, F. F. **Mineração de dados aplicada à análse do comportamento de válvulas gaveta submarinas na indústria de óleo e gás**. Monografia—Macaé - RJ: Faculdade Salesiana Maria Auxiliadora, 2019.

NADEMI, H.; VANFRETTI, L.; PRETLOVE, J. Fault detection method in subsea power distribution systems using statistical optimisation. **IET Energy Systems Integration**, v. 2, n. 2, p. 144–150, 1 jun. 2020.

PONTIUS, R. G.; PARMENTIER, B. Recommendations for using the relative operating characteristic (ROC). **Landscape Ecology**, v. 29, n. 3, p. 367–382, mar. 2014.

SHYU, M.-L. et al. A Novel Anomaly Detection Scheme Based on Principal Component Classifier. p. 10, [s.d.].

Apêndice A – Código

Ver anexos do PDF ou Código no GitHub.