DOI: 10.48011/asba.v2i1.1405

Detecção de anomalias em poços produtores de petróleo usando aprendizado de máquina

Wander Fernandes Júnior [*] **Ricardo Emanuel Vaz Vargas** [**]
Karin Satie Komati [*] **Kelly Assis de Souza Gazolli** [*]

[*] *Programa de Pós-Graduação em Computação Aplicada (PPComp) Instituto Federal do Espírito Santo, Rodovia ES-010 - Km 6,5 - Manguinhos, Serra - ES, CEP 29173-087.*
[**] *Petróleo Brasileiro S.A., Avenida Nossa Senhora da Penha - 1688 - Barro Vermelho, Vitória - ES, CEP 29057-570.*

E-mails: wanderfj@gmail.com, ricardo.vargas@petrobras.com.br kkomati@ifes.edu.br, kasouza@ifes.edu.br,

Abstract: Anomalies in oil production wells can cause significant financial impacts. The use of machine learning to detect these situations can prevent unwanted production interruptions as well as maintenance costs. In this context, this work proposes the application and comparison of classifiers for detecting anomalies in oil and gas production wells. One-class classifiers such as Isolation Forest, One-class Support Vector Machine (OCSVM), Local Outlier Factor (LOF), and Elliptical Envelope were applied to a dataset with real cases. The best performance was obtained by LOF with an F1-score of 88.2%, followed by Isolation Forest with 74.3%. The results obtained show improvement in comparison to the reference benchmark and stimulate the continuation of work with experimentation of other families of classifiers.

Resumo: Anomalias em poços produtores de petróleo podem provocar impactos financeiros significativos. O uso de aprendizado de máquina para detectar essas situações podem prevenir interrupções indesejadas de produção bem como custos de manutenção. Nesse contexto, este trabalho propõe a aplicação e comparação de classificadores para detecção de anomalias em poços de produção de petróleo e gás. Classificadores de classe única Floresta de Isolamento, *One-class Support Vector Machine* (OCSVM), *Local Outlier Factor* (LOF) e Envelope Elíptico foram aplicados em uma base de dados com casos reais, sendo o melhor desempenho obtido pelo LOF com medida F1 de 88,2%, seguido da Floresta de Isolamento com 74,3%. Os resultados obtidos apresentam melhoria em comparação ao *benchmark* de referência e estimulam a continuação do trabalho com a experimentação de outras famílias de classificadores.

Keywords: Anomaly Detection; Oil Well Monitoring; Multivariate Time Series.

Palavras-chaves: Detecção de Anomalias; Monitoramento de Poços de Petróleo; Séries Temporais Multivariadas.

1. INTRODUÇÃO

Petróleo é uma matéria-prima essencial à vida moderna, sendo componente básico para diversos tipos de indústrias. Dele se produz gasolina, combustível de aviação, gás de cozinha, lubrificantes, borrachas, plásticos, tecidos sintéticos, tintas e energia elétrica (Gauto et al., 2016).

Conforme dados da Agência Nacional do Petróleo, Gás Natural e Biocombustíveis (ANP, 2020), a produção de petróleo e gás natural no Brasil em janeiro de 2020 foi de 3,168 MMbbl/d (milhões de barris por dia) e 139 MMm^3/d (milhões de metros cúbicos por dia), respectivamente. Essa produção foi proveniente de 7.227 poços, sendo 649 marítimos e 6.558 terrestres. A produção do pré-sal [1] correspondeu a 66,4% desse total e foi oriunda de 119 poços marítimos, o que equivale a uma média de produção de aproximadamente 18 Mbbl/d (milhares de barris por dia).

Durante a produção de petróleo é possível a ocorrência de eventos indesejados denominados anomalias, que podem provocar impactos financeiros significativos. Como exemplo, pode-se citar a incrustação (ocorrência de depósitos inorgânicos) em válvulas que podem reduzir drasticamente a produção de petróleo (Vargas, 2019).

Assim, é importante que o processo de produção de petróleo seja monitorado a fim de detectar e classificar anomalias. De acordo com (Qin, 2012), uma possível solução é a aplicação de estatísticas multivariadas e métodos de aprendizado de máquina para detecção e classificação de anomalias.

Na área de aprendizado de máquina, o problema de classificação pode ser definido como a categorização de uma

[1] "Pré-sal" refere-se à produção de hidrocarbonetos realizada no horizonte geológico denominado "Pré-sal", em campos localizados na área definida no inciso IV do caput do art. 2º da Lei nº 12.351, de 2010.

determinada entrada em uma ou mais classes discretas e pré-definidas (Kadhim, 2019). Em muitos processos industriais, busca-se detectar padrões raros, nos quais a maioria das observações referem-se a situações de normalidade e a minoria, às situações raras que se deseja identificar (Santos and Kern, 2016). Nesses casos a detecção de padrões novos (*novelty detection*) pode ser feita com classificadores de classe única, nos quais utiliza-se no treinamento apenas dados associados a classe comum (normalidade) (Khan and Madden, 2014).

Em processos industriais, os dados de entrada para o monitoramento são provenientes de sensores e indexados no tempo, ou seja, são séries temporais multivariadas. Conforme escrito por Fawaz et al. (2019), nas duas últimas décadas a classificação em séries temporais tem sido considerada como um dos problemas mais desafiadores em mineração de dados. Uma das dificuldades é que as anomalias não possuem um conjunto de características ou regras que as agregam. Um anomalia pode ser pontual, isto é, um único valor extremo (como um valor de temperatura acima de limiar) pode ser o suficiente para caracterizar uma anomalia. Mas um valor que pontualmente pode ser considerado normal pode ser considerado anormal em um determinado contexto, como por exemplo, uma mudança brusca de temperatura durante um processo industrial, mesmo que o valor inicial e final da mudança não sejam atípicos isoladamente (Chandola et al., 2009). Além disso, o que é atípico em um sistema pode não ser atípico em outro sistema, pois cada um pode ter suas individualidades.

Este trabalho se propõe a aplicar e comparar técnicas de aprendizado de máquina na detecção de anomalias em poços produtores de petróleo, utilizando a base de dados 3W (Vargas et al., 2019), composta por séries temporais multivariadas. As técnicas de classificadores de classe única comparadas são: Floresta de Isolamento *Isolation Forest*, *One-class Support Vector Machine* (OCSVM), *Local Outlier Factor* (LOF) e Envelope Elíptico. Não fez parte do escopo classificar o tipo da anomalia.

A base de dados 3W (Vargas et al., 2019) é pública e contém 1.984 instâncias de séries temporais da produção de poços de petróleo. Essas instâncias foram separadas em: operação em condições normais e anomalias. As anomalias foram organizadas em oito classes. Essa base pode ser utilizada tanto para detecção quanto para classificação de anomalias em poços de petróleo.

No que se refere à estrutura do artigo, inicia-se com uma seção de apresentação de trabalhos correlatos e recentes sobre detecção de anomalias. Em seguida, são apresentados os principais conceitos relativos a sistemas de produção de petróleo e a descrição da base de dados utilizada. A abordagem proposta é descrita na seção subsequente, formando a base para o detalhamento dos experimentos e para a apresentação e discussão dos resultados obtidos.

2. TRABALHOS CORRELATOS

O trabalho de Chandola et al. (2009) é um importante artigo *survey* no tema. Embora não seja recente, traz contribuições e discussões sobre o conceito de anomalia, seus diferentes aspectos em cada domínio de aplicação, dando uma visão geral estruturada, agrupando técnicas existentes em diferentes categorias, identificando as vantagens e desvantagens de cada uma. Também fornece uma discussão sobre a complexidade computacional das técnicas.

Barbariol et al. (2019) propuseram uma abordagem de detecção de anomalias em módulos de metrologia de medidores de fluxo multifásicos. Esses equipamentos são importantes ferramentas no setor de petróleo e gás, pois fornecem simultaneamente dados em tempo real dos fluxos de óleo, gás e água. Os algoritmos *Cluster Based Local Outlier Factor* e Floresta de Isolamento foram utilizados para detectar alterações de qualidade nas medições realizadas, tendo sido utilizado um conjuntos de dados semi-sintéticos.

Chan et al. (2019) realizaram detecção de anomalias em controladores lógicos programáveis (CLPs) que compõem sistemas de controle de supervisão e aquisição de dados (SCADA). Esses equipamentos gerenciam operações de equipamentos industriais baseados em sensores e estão expostos a ameaças cibernéticas. Foi realizado um estudo de caso envolvendo uma simulação de semáforo que demonstrou que as anomalias são detectadas com alta precisão utilizando *One-class* SVM.

Khan et al. (2019) aplicaram técnicas de detecção de anomalias em veículos aéreos não tripulados. Foram utilizados dados de uma base denominada *Aero-Propulsion System Simulation* e realizados experimentos em um veículo real. Foram investigados os requisitos para aplicativos de engenharia e demonstrada uma implementação do algoritmo de Floresta de Isolamento.

Tan et al. (2020) compararam o desempenho de diversos classificadores para detecção de anomalias em máquinas de embarcações marítimas. A segurança e a confiabilidade da navegação dependem do desempenho dessas máquinas e o monitoramento inteligente de condições é importante para as atividade de manutenção. Um conjunto de dados de um sistema de propulsão de turbina a gás de um navio foi utilizado. Foi investigado o desempenho de classificadores de classe única: OCSVM, SVDD (*Support Vector Data Description*), GKNN (*Global K-Nearest Neighbors*), LOF, Floresta de Isolamento e ABOD (*Angle-based Outlier Detection*). Os resultados mostraram bom desempenho dos algoritmos no conjunto de dados utilizado.

Os trabalhos de Vargas et al. (2019) e Vargas (2019) são a base deste trabalho. Eles elaboraram e tornaram público a base 3W e usaram as técnicas de Floresta de Isolamento e OCSVM para detecção de anomalias. A proposta deste trabalho estende o de Vargas (2019), pois usa mais classificadores: o LOC e o Envelope Elíptico, além de adicionar o passo de calibração dos hiperparâmetros das técnicas de Floresta de Isolamento e OCSVM (que não foi realizada no trabalho base).

3. PRODUÇÃO DE PETRÓLEO E A BASE 3W

Nas subseções a seguir, são apresentados os principais conceitos sobre sistemas de produção de petróleo, anomalias em poços de petróleo e uma descrição mais detalhada da base 3W.

Um poço de petróleo é uma estrutura perfurada no solo em etapas que formam um telescópio invertido (os diâmetros diminuem à medida que a profundidade aumenta) e mu-

nida com equipamentos e sensores que permitem o fluxo de petróleo e gás da rocha reservatório de petróleo até a superfície (Guo, 2011).

Para que a produção de petróleo e gás seja possível no ambiente marítimo, os poços são conectados a sistemas compostos de equipamentos submarinos instalados no leito marinho e linhas que permitem o controle do poço e o escoamento do petróleo até uma plataforma de produção, armazenamento e transferência (Bai and Bai, 2015).

A Figura 1 traz um esquema simplificado de um sistema de produção de petróleo, contemplando o poço, o sistema submarino e a plataforma. O óleo e o gás fluem de uma rocha reservatório de petróleo através da tubulação de produção e, em seguida, através de uma linha de produção para uma plataforma.

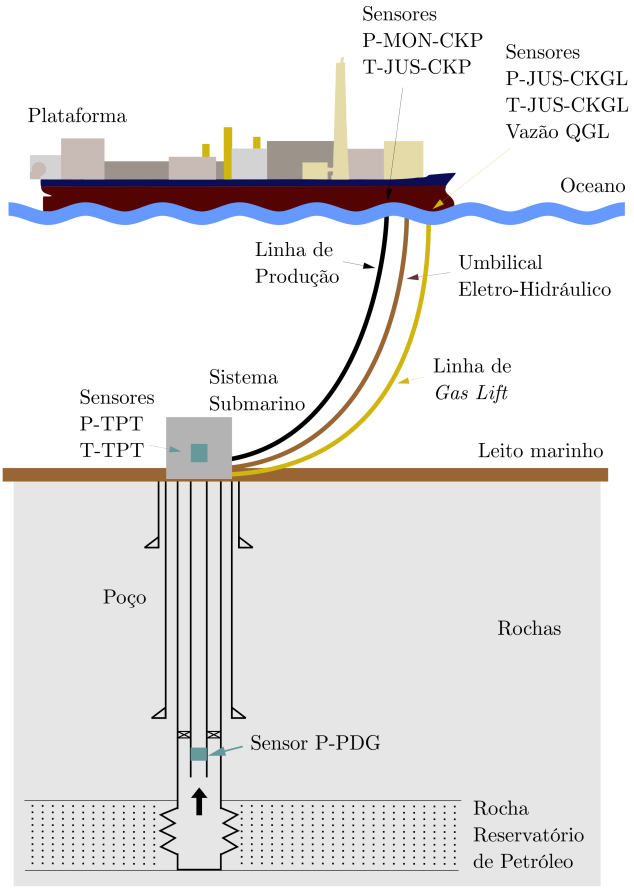

Figura 1. Esquema simplificado de um sistema de produção de petróleo. Fonte: elaborado pelo próprio autor.

As válvulas instaladas no fundo do mar são operadas remotamente por um umbilical eletro-hidráulico. Existem dispositivos sensores que auxiliam no monitoramento: um manômetro permanente de fundo de poço (P-PDG), um transdutor de temperatura (T-TPT) e um transdutor de pressão (P-TPT).

A DHSV (*Down Hole Safety Valve*) é uma válvula de segurança instalada na tubulação de produção de poços. Seu objetivo é garantir o fechamento do poço no caso de uma situação em que a unidade de produção e o poço estejam fisicamente desconectados ou no caso de uma emergência ou falha catastrófica do equipamento de superfície. A válvula CKP (*Choke* de Produção) localiza-se na plataforma e é responsável pelo controle da abertura do poço, possui sensores de temperatura (T-JUS-CKP) e de pressão (P-MON-CKP). A linha de *gas lift* na plataforma tem sensores de vazão (vazão QGL), temperatura (T-JUS-CKGL) e pressão (P-JUS-CKGL).

3.1 Base 3W

A base de dados utilizada neste trabalho foi publicada por Vargas et al. (2019) e é intitulada 3W. Cada instância é composta por oito variáveis (oito séries temporais), conforme descrito na Tabela 1, provenientes de sensores de sistemas de produção de petróleo conforme a localização física aproximada mostrada na Figura 1. A cada instância existe uma variável adicional que é um vetor de rótulos no nível de observação que estabelece até três períodos em cada instância de qualquer tipo: normal, transiente de anomalia e estado estável de anomalia.

Tabela 1. Descrição das séries temporais.

Variável	Descrição
P-PDG	Pressão do fluido no PDG
P-TPT	Pressão do fluido no TPT
T-TPT	Temperatura do fluido no TPT
P-MON-CKP	Pressão do fluido montante à válvula CKP
T-JUS-CKP	Temperatura do fluido jusante à válvula CKP
P-JUS-CKGL	Pressão do fluido jusante à válvula de controle de *gas lift*
T-JUS-CKGL	Temperatura do fluido jusante à válvula de controle de *gas lift*
QGL	Vazão de *gas lift*.
Vetor de Rótulo	Indica o estado de cada anomalia ao longo da série temporal: período normal, transiente de anomalia e estado estável de anomalia.

Do total de 1.984 instâncias, são 597 normais e as demais (1.397) são anomalias. A base 3W categoriza as anomalias em oito principais classes, que são listadas a seguir com a indicação da quantidade de instâncias de cada anomalia.

(1) Aumento Abrupto de BSW (129 instâncias): o *Basic Sediment and Water* (BSW) é definido como a razão entre a produção de água e a produção total (óleo+água) do poço. Um aumento abrupto desse valor pode acarretar dificuldades de escoamento e elevação, menor produção de óleo e menor fator de recuperação de petróleo.
(2) Fechamento Espúrio de DHSV (38 instâncias): o fechamento espúrio dessa válvula causa paradas de produção.
(3) Intermitência Severa (106 instâncias): condição de instabilidade de escoamento de grande duração, grande amplitude e periodicidade definida que causa redução da produção e danos às instalações.
(4) Instabilidade de Fluxo (344 instâncias): comportamento instável de escoamento (também chamado de golfadas) que ocorre com frequência em poços de produção de petróleo e gás.
(5) Perda Rápida de Produtividade (451 instâncias): alteração de propriedades do reservatório de petróleo (pressão, razão gás/óleo) ou do fluido produzido (den-

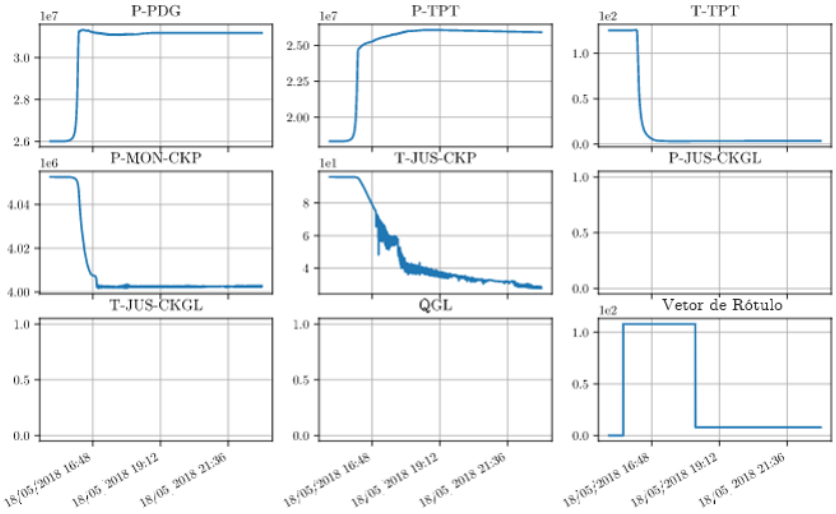

Figura 2. Exemplo de instância da classe “Aumento Abrupto de BSW” da base de dados 3W.

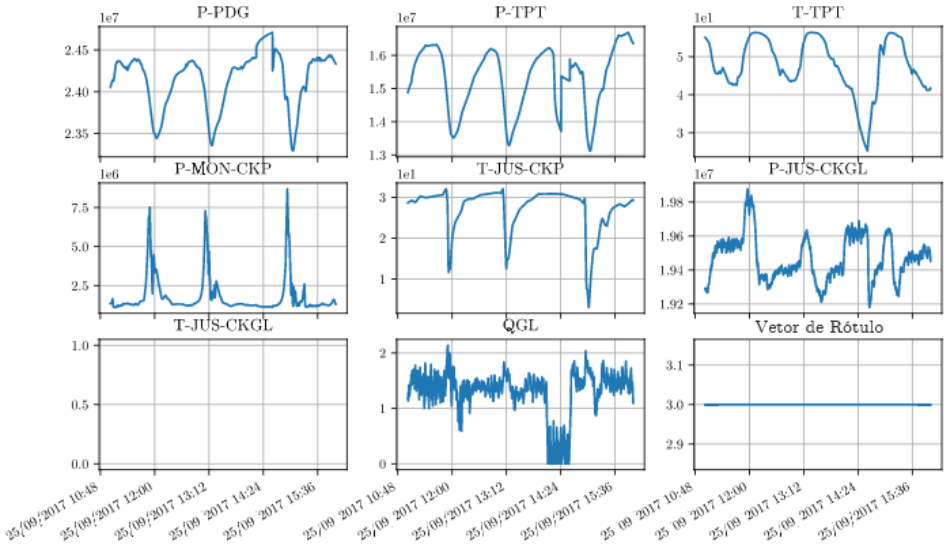

Figura 3. Exemplo de instância da classe “Intermitência Severa” da base de dados 3W.

sidade, viscosidade) que dificulta o escoamento do petróleo.

(6) Restrição Rápida em CKP (221 instâncias): eventuais restrições rápidas e indesejadas podem ocorrer nessa válvula por problemas operacionais.
(7) Incrustação em CKP (14 instâncias): ocorrência de depósitos inorgânicos na válvula CKP.
(8) Hidrato em Linha de Produção (84 instâncias): hidrato é um composto cristalino sólido formado por água e gás natural (assemelha-se ao gelo comum).

A Figura 2 apresenta exemplo de instância da classe “Aumento Abrupto de BSW”. No gráfico “Vetor de rótulo”, é possível verificar que inicialmente estava no valor 0,0 (classe normal). Em seguida, a instância passa pelo transiente de anomalia (vários outros gráficos apresentam mudança abrupta de valores) até que atinge o estado estável de anomalia no valor 1,0 (numeração associada à classe em questão). Em três dos gráficos não há sinais, pois são valores ausentes na base.

A base de dados tem 4.947 variáveis ausentes (valores indisponíveis por problemas em sensores ou redes de comunicação ou por inaplicabilidade da variável à instância) que representam 31,17% de todas as 15.872 variáveis de todas as 1.984 instâncias. Também tem 1.535 variáveis congeladas (valores que se mantêm fixos por conta de problemas em sensores ou redes de comunicação) que representam 9,67% de todas as 15.872 variáveis de todas as 1.984 instâncias.

A Figura 3 apresenta exemplo da instância de classe “Intermitência Severa”. No gráfico “Vetor de rótulo”, o valor é constante em 3,0 (numeração associada à classe em questão) no estado estável de anomalia. Os períodos normal e transiente de anomalia não existem nesse exemplo. Nos

gráficos é possível verificar que há oscilação dos valores de pressão, temperatura e vazão nos vários sensores, demonstrando irregularidades descontinuadas.

Um arquivo específico e padronizado com formato *Comma-Separated Values* (CSV) foi gerado para cada instância. Esses arquivos foram agrupados em diretórios baseados na classe da anomalia. Todas as instâncias foram geradas com observações obtidas com taxa de amostragem fixa (1 Hz) e a origem de cada instância (real, simulada ou desenhada) foi incorporada ao nome do arquivo.

4. ABORDAGEM PROPOSTA

Nesta seção apresentam-se os métodos empregados na proposta deste trabalho. Um processo de classificação pode ser dividido em cinco etapas: coleta de dados, pré-processamento, extração de características, aplicação da técnica de classificação e avaliação de desempenho (Kadhim, 2019).

A coleta de dados é a fase inicial, na qual é realizado o levantamento dos dados a serem utilizados, que neste caso é a base 3W. As subseções a seguir descrevem cada uma das etapas após a coleta de dados.

4.1 Pré-processamento

A etapa de pré-processamento trata da preparação inicial do conjunto de dados coletado. No caso de séries temporais, nessa etapa inclui-se a análise dos dados, geração de gráficos para entendimento do dados, remoção de valores nulos e/ou congelados, e reamostragens de observações das séries temporais para balanceamento da base de dados (Pal and Prakash, 2017).

Na etapa de pré-processamento foi realizada amostragem das instâncias com janela deslizante com geração de até 15 amostras com 180 observações cada. Dos períodos normais as primeiras observações foram utilizadas para treinamento (60%) e as últimas, para teste (40%). Dos períodos rotulados como anomalias, as observações foram utilizadas apenas para teste.

As variáveis das amostras foram normalizadas de forma que todas elas passaram a ter média zero e variância unitária. As variáveis das amostras utilizadas no treinamento que tinham quantidades de valores ausentes acima de um limiar (10%) ou que tinham desvios padrões abaixo de outro limiar (1%) foram totalmente descartadas.

4.2 Extração de características

Em seguida, na etapa de extração de características, os dados pré-processados são trabalhados para obter as características mais relevantes para a classificação.

Uma série temporal univariada $x = [x_1, x_2, ..., x_T]$ é um conjunto ordenado de valores reais. O comprimento de x é igual ao número de valores reais T. A série temporal é multivariável $X = [X^1, X^2, ..., X^M]$ quando consiste em M séries temporais univariadas diferentes com $X^i \in R^T$ (Fawaz et al., 2019).

A partir de cada amostra de séries temporal, foram extraídas e utilizadas como características a mediana, média, desvio padrão, variância, máximo e mínimo para cada variável. Para esta extração de características das séries temporais foi utilizada a biblioteca $tsfresh$ [2] (*Time series feature extraction*) (Christ et al., 2018).

4.3 Classificação

Classificadores de classe única podem ser utilizados na detecção de padrões raros. Em geral utiliza-se apenas a classe comum (normalidade) no treinamento (*novelty detection*) e nos testes há uma mistura de instâncias normais e anormais (Khan and Madden, 2014). Foram escolhidas as seguintes técnicas: *One-class SVM*, Floresta de Isolamento, *Local Outlier Factor* (LOF) e Envelope Elíptico, que serão explicadas a seguir.

***One-class Support Vector Machine** (OCSVM)* Baseado em otimização, o SVM constrói hiperplanos para separar diferentes classes no espaço. O objetivo é maximizar o parâmetro b (chamado de margem) que representa a distância entre o hiperplano e o primeiro ponto de cada classe de forma a criar um limiar de decisão (Duda et al., 2012).

Existem versões alternativas que utilizam funções matemáticas não lineares (chamadas *kernels*) que mapeiam o espaço de características em outro de maior dimensionalidade no qual a separabilidade entre as classes tende a ser maior (Géron, 2019).

O SVM de classe única foi introduzido por Schölkopf et al. (2001) e é ilustrado na Figura 4. Tem como objetivo construir uma hiperesfera que engloba todas as instâncias normais em um espaço. Uma nova instância é classificada como uma anomalia quando não se enquadra no espaço dessa hiperesfera (Misra et al., 2020).

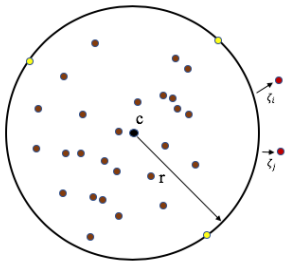

Figura 4. A hiperesfera que contém os dados normais com centro 'c' e raio 'r'. Os objetos no limite (em amarelo) são vetores de suporte e dois objetos ficam fora do limite são consideradas as anormalidades. Fonte: imagem GNU *Free Licence*.

É possível ajustar os hiperparâmetros de *kernel* (linear, polinomial, radial), *gamma* (gama) e *nu* (ni). O parâmetro *gama* influencia o raio da hiperesfera gaussiana que separa as instâncias normais das anomalias - grandes valores de gama resultam em uma hiperesfera menor e em um modelo "mais rígido" que encontra mais discrepâncias. A fração *ni* define a porcentagem do conjunto de dados que é discrepante e ajuda a criar limites de decisão mais rígidos (Misra et al., 2020).

[2] https://tsfresh.readthedocs.io/

Floresta de Isolamento Baseada em busca, árvores de decisão são construídas com base em regras inferidas a partir dos atributos. Embora não tenham sido projetadas originalmente para o problema de detecção de anomalias, é possível sua utilização por meio da análise dos caminhos percorridos na árvore durante a divisão em cada nó (Aggarwal and Sathe, 2017).

Liu et al. (2012) denominaram árvore de isolamento quando em cada nó um atributo é selecionado aleatoriamente e, em seguida, divide-se o conjunto de dados em dois a partir de um valor limite aleatório (entre os valores mínimo e máximo). O conjunto de dados é gradualmente dividido até que todas as instâncias sejam isoladas (Géron, 2019). Como anomalias em geral estão em regiões menos populosas do conjunto de dados, geralmente são necessárias menos partições aleatórias para isolá-las em nós da árvore (Misra et al., 2020).

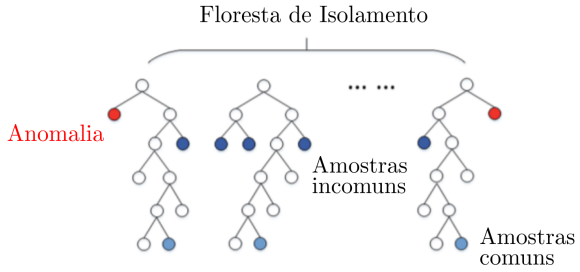

Figura 5. Floresta de isolamento, onde os nós vermelhos são as anomalias. Fonte: adaptado de Chen et al. (2016).

Uma técnica é chamada de *ensemble* quando um conjunto de classificadores é treinado individualmente mas as decisões são tomadas de forma combinada. Métodos *ensemble* tendem a apresentar um menor *overfitting* (Aggarwal and Sathe, 2017). O método *ensemble* Floresta de Isolamento busca criar uma estrutura de árvores aleatórias para isolar as anomalias das instâncias. Conforme ilustrado na Figura 5, as anomalias tem maior suscetibilidade ao isolamento e ficam mais perto das raízes das árvores, enquanto os pontos normais são difíceis de isolar e geralmente estão no extremo mais profundo da árvore. Os comprimentos médios de caminho em várias árvores são utilizados para obter uma pontuação e classificar a instância (Chen et al., 2016).

Na técnica de Floresta de Isolamento é possível controlar o desempenho por meio de ajuste dos seguintes hiperparâmetros: número de estimadores, número máximo de amostras e características utilizados por cada árvore, reamostragem e contaminação no conjunto de dados (Misra et al., 2020).

***Local Outlier Factor** (LOF)* Baseado em densidade, o LOF foi desenvolvido por Breunig et al. (2000) e compara a densidade de instâncias em torno de uma determinada instância com a densidade em torno de seus vizinhos. Uma anomalia geralmente é mais isolada que seus vizinhos mais próximos.

O algoritmo calcula pontuação LOF de uma observação como a razão entre a densidade local média de seus k vizinhos mais próximos e sua própria densidade local. Ao comparar a densidade local de uma amostra com as densidades locais de seus vizinhos, pode-se identificar amostras que possuem uma densidade substancialmente menor do que seus vizinhos. São consideradas como instâncias normais as que têm densidade local semelhante a de seus vizinhos, enquanto são consideradas anomalias as que têm densidade local menor (Misra et al., 2020). Ideia ilustrada na Figura 6.

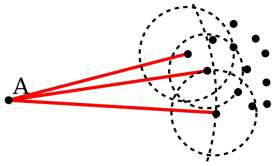

Figura 6. Ideia básica do LOF: comparando a densidade local de um ponto com as densidades de seus vizinhos. O elemento 'A' tem uma densidade muito menor do que seus vizinhos. Imagem com permissão de uso de domínio público.

No LOF é possível ajustar os hiperparâmetros: número k de vizinhos a serem considerados, tamanhos de folha do algoritmo, métrica de distância e contaminação no conjunto de dados. O parâmetro $novelty = True$ permite que a técnica seja aplicada em novos dados.

Envelope Elíptico O Envelope Elíptico implementa a técnica *Minimum Covariance Determinant* (MCD) que assume que as instâncias normais são geradas a partir de uma única distribuição gaussiana e que o conjunto de dados está contaminado com anomalias que não foram geradas a partir dessa distribuição (Géron, 2019).

O algoritmo estima os parâmetros da distribuição gaussiana (forma do envelope elíptico ao redor das instâncias normais) e fornece uma estimativa de um envelope elíptico que permite a identificação das anomalias (Géron, 2019), tal como ilustrado na Figura 7.

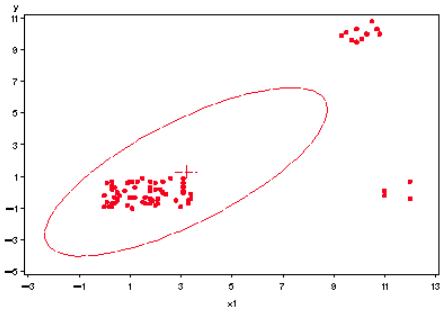

Figura 7. Envelope elíptico em que todos os elementos internos à figura geométrica são normais e todos os elementos externos à elipse são anormalidades.

O Envelope Elíptico permite o ajuste dos parâmetros de centralização, fração de suporte e contaminação no conjunto de dados.

4.4 Avaliação de Desempenho

Na última etapa do processo de classificação, é necessário avaliar o desempenho obtido. Essa avaliação é feita por

meio de métricas obtidas a partir de uma matriz de confusão, tais como acurácia, revocação, precisão e/ou medida F1. Nesse trabalho utilizou-se a medida F1 para comparação de resultados com o trabalho de (Vargas, 2019).

5. EXPERIMENTOS, RESULTADOS E DISCUSSÃO

Os experimentos foram realizados seguindo-se as seguintes regras estabelecidas no *benchmark* de detecção de anomalia:

- Apenas instâncias reais com anomalias de tipos que têm períodos normais maiores ou iguais a vinte minutos foram utilizadas;
- Múltiplas rodadas de treinamento e teste foram realizadas, sendo o número de rodadas igual ao número de instâncias. Em cada rodada, as amostras utilizadas para treinamento ou teste foram extraídas de apenas uma instância. Parte das amostras de normalidade foram utilizadas no treinamento e a outra parte, no teste. Todas as amostras de anormalidades foram utilizadas apenas no teste (técnica de aprendizagem de classe única). O conjunto de teste foi composto pelo mesmo número de amostras de cada classe (normalidade e anormalidade);
- Em cada rodada, precisão, revocação e medida F1 foram computadas (valor médio e desvio padrão de cada métrica), sendo o valor médio da medida F1 apresentada nesta seção para comparação com trabalho anterior (Vargas, 2019).

5.1 Calibração dos classificadores

A calibração dos classificadores (implementado pela função *ParameterGrid* do scikit-learn) gerou 1.716 diferentes combinações entre classificadores e hiperparâmetros. Na Tabela 2 constam os hiperparâmetros que foram utilizados para obtenção das métricas exibidas na Tabela 3.

Tabela 2. Hiperparâmetros utilizados para obtenção das métricas apresentadas.

Classificador	Parâmetro	Valor
Local Outlier Factor	n_neighbors	10
	leaf_size	15
	metric	'chebyshev'
	contamination	'auto'
Floresta de Isolamento	n_estimators	150
	max_samples	1.0
	max_features	1.0
	bootstrap	False
	contamination	0
Envelope Elíptico	assume_centered	False
	support_fraction	0.8
	contamination	0.0001
One-class SVM	kernel	'rbf'
	gamma	0.001
	nu	0.1

5.2 Resultados dos classificadores

Na Tabela 3 constam os melhores resultados de medida F1 e desvio padrão obtidos na aplicação dos classificadores às instâncias da base de dados 3W. Todos os resultados deste trabalho são reprodutíveis, pois o código fonte utilizado encontra-se disponível em um repositório GitHub [3]. Um classificador ingênuo (*dummy*) que classifica todas as instâncias do conjunto de teste como normais foi incluído para aprimorar a comparação de resultados, suas métricas podem ser usadas como um valor mínimo de referência.

Tabela 3. Medida F1 e desvio padrão dos classificadores experimentados.

Classificador	F1 (Média)	Desvio Padrão
Local Outlier Factor	0,882	0,126
Floresta de Isolamento	0,743	0,179
Envelope Elíptico	0,664	0,157
One-class SVM	0,567	0,162
Dummy (ingênuo)	0,500	0,000

O classificador que obteve a melhor medida F1 nos testes foi o LOF com medida F1 de 0,882 seguido da Floresta de Isolamento com F1 de 0,743. Presume-se que pelo fato do LOF ter apresentado melhor resultado significa que a definição de fronteiras dos casos normais em uma única classe, como no OCSVM e no Envelope Elíptico que apresentaram os menores valores, não é bem definido. E que mesmo os casos normais são melhor representados por vários grupamentos, e por isso a medida F1 foi maior para a Floresta de Isolamento e para o LOF.

Os resultados obtidos apresentam melhorias em relação aos resultados obtidos por Vargas (2019) e mostrados na Tabela 4 (os classificadores *Local Outlier Factor* e Envelope Elíptico não foram utilizados no trabalho de referência). A medida F1 foi 0,727 para Floresta de Isolamento e 0,532 para *One-class SVM*, respectivamente. A calibração de cada classificador melhorou um pouco a métrica medida F1 e diminuiu o desvio padrão.

Tabela 4. Medida F1 e desvio padrão do *benchmark* de referência (resultados anteriores).

Classificador	F1 (Média)	Desvio Padrão
Floresta de Isolamento	0,727	0,182
One-class SVM	0,532	0,075

Os testes estatísticos (não-paramétricos) utilizados em Vargas (2019) e descritos em Demšar (2006) também foram utilizados neste trabalho. Os seus resultados foram analisados considerando-se significância de 5%.

A verificação se múltiplos classificadores podem gerar métricas F1 cujas médias sejam iguais entre si foi feita com o Teste de Friedman. Como o resultado desse teste (valor $p = 1,110 \times 10^{-16}$) rejeitou a hipótese nula, pôde-se concluir que ao menos um dos classificadores testados gera métricas F1 cuja média é diferente em relação às demais com alta probabilidade.

Na sequência, a verificação de quais classificadores baseados nos quatro algoritmos considerados podem geram métricas F1 cujas médias sejam iguais à média obtida pelo classificador ingênuo foi feita com o Teste de Holm. Nesse teste, utilizou-se as mesmas métricas F1 submetidas ao Teste de Friedman, a Correção de Bonferroni e o classificador ingênuo como grupo de controle. Os resultados obtidos são apresentados na Tabela 5 e mostram que a hipótese

[3] https://github.com/wanderfernandesjunior/anomaly-detection

nula não pode ser rejeitada apenas para o classificador baseado em *One-Class SVM*.

Tabela 5. Resultados do Teste de Holm.

Classificador	Valor p	Valor p corrigido
Local Outlier Factor	0,000	0,000
Floresta de Isolamento	$2,469 \times 10^{-8}$	$3,814 \times 10^{-5}$
Envelope Elíptico	$7,794 \times 10^{-6}$	$9,571 \times 10^{-3}$
One-class SVM	0,168	1,000

Portanto, em função dos resultados apresentados nas Tabelas 3 e 5, pôde-se concluir que os classificadores baseados em *Local Outlier Factor*, Floresta de Isolamento e Envelope Elíptico geram, com alta probabilidade, métricas F1 cujas médias são diferentes e maiores em relação à média de métricas F1 obtidas com o classificador ingênuo.

6. CONCLUSÃO

Este trabalho realizou uma comparação das técnicas de detecção de anomalias em poços de produção de petróleo utilizando classificadores de classe única.

Foi utilizado como referência o *benchmark* proposto por Vargas et al. (2019) no qual demonstrou-se ser possível detectar anomalias em poços com aprendizado de máquina. Os resultados obtidos apresentaram melhoria em relação ao *benchmark* e estimulam a continuação do trabalho com a experimentação de outras técnicas - tanto para extração de características quanto para utilização de outras famílias de algoritmos para detecção de anomalias, tais como os baseados em redes neurais.

REFERÊNCIAS

Aggarwal, C.C. and Sathe, S. (2017). *Outlier ensembles: An introduction*. Springer.

ANP (2020). Boletim mensal da produção de petróleo e gás natural. URL `http://www.anp.gov.br/`. [acessado em 25-03-2020].

Bai, Q. and Bai, Y. (2015). *Sistemas marítimos de produção de petróleo: processos, tecnologias e equipamentos offshore*. Elsevier Brasil.

Barbariol, T., Feltresi, E., and Susto, G.A. (2019). Machine learning approaches for anomaly detection in multiphase flow meters. *IFAC-PapersOnLine*, 52(11), 212–217.

Breunig, M.M., Kriegel, H.P., Ng, R.T., and Sander, J. (2000). Lof: identifying density-based local outliers. In *Proceedings of the 2000 ACM SIGMOD international conference on Management of data*, 93–104.

Chan, C.F., Chow, K.P., Mak, C., and Chan, R. (2019). Detecting anomalies in programmable logic controllers using unsupervised machine learning. In *IFIP International Conference on Digital Forensics*, 119–130. Springer.

Chandola, V., Banerjee, A., and Kumar, V. (2009). Anomaly detection: A survey. *ACM Computing Surveys*, 41(3), 1–58. doi:10.1145/1541880.1541882.

Chen, W.R., Yun, Y.H., Wen, M., Lu, H.M., Zhang, Z.M., and Liang, Y.Z. (2016). Representative subset selection and outlier detection via isolation forest. *Analytical methods*, 8(39), 7225–7231.

Christ, M., Braun, N., Neuffer, J., and Kempa-Liehr, A.W. (2018). Time series feature extraction on basis of scalable hypothesis tests (tsfresh–a python package). *Neurocomputing*, 307, 72–77.

Demšar, J. (2006). Statistical comparisons of classifiers over multiple data sets. *Journal of Machine learning research*, 7(Jan), 1–30.

Duda, R.O., Hart, P.E., and Stork, D.G. (2012). *Pattern classification*. John Wiley & Sons.

Fawaz, H.I., Forestier, G., Weber, J., Idoumghar, L., and Muller, P.A. (2019). Deep learning for time series classification: a review. *Data Mining and Knowledge Discovery*, 33(4), 917–963.

Gauto, M.A., de Melo Apoluceno, D., Amaral, M.C., and Auríquio, P.C. (2016). *Petróleo e gás: princípios de exploração, produção e refino*. Bookman Editora.

Géron, A. (2019). *Hands-on Machine Learning with Scikit-Learn, Keras, and TensorFlow*. "O'Reilly Media, Inc.".

Guo, B. (2011). *Petroleum production engineering, a computer-assisted approach*. Elsevier.

Kadhim, A.I. (2019). Survey on supervised machine learning techniques for automatic text classification. *Artificial Intelligence Review*, 1–20.

Khan, S., Liew, C.F., Yairi, T., and McWilliam, R. (2019). Unsupervised anomaly detection in unmanned aerial vehicles. *Applied Soft Computing*, 83, 105650.

Khan, S.S. and Madden, M.G. (2014). One-class classification: taxonomy of study and review of techniques. *The Knowledge Engineering Review*, 29(3), 345–374.

Liu, F.T., Ting, K.M., and Zhou, Z.H. (2012). Isolation-based anomaly detection. *ACM Transactions on Knowledge Discovery from Data (TKDD)*, 6(1), 1–39.

Misra, S., Li, H., and He, J. (2020). *Machine Learning for Subsurface Characterization*. Elsevier.

Pal, A. and Prakash, P. (2017). *Practical Time Series Analysis: Master Time Series Data Processing, Visualization, and Modeling using Python*. Packt Publishing Ltd.

Qin, S.J. (2012). Survey on data-driven industrial process monitoring and diagnosis. *Annual reviews in control*, 36(2), 220–234.

Santos, T. and Kern, R. (2016). A literature survey of early time series classification and deep learning. In *Sami@ iknow*.

Schölkopf, B., Platt, J.C., Shawe-Taylor, J., Smola, A.J., and Williamson, R.C. (2001). Estimating the support of a high-dimensional distribution. *Neural computation*, 13(7), 1443–1471.

Tan, Y., Tian, H., Jiang, R., Lin, Y., and Zhang, J. (2020). A comparative investigation of data-driven approaches based on one-class classifiers for condition monitoring of marine machinery system. *Ocean Engineering*, 201, 107174.

Vargas, R.E.V. (2019). *Base de Dados e Benchmarks para Prognóstico de Anomalias em Sistemas de Elevação de Petróleo*. Tese de doutorado, Universidade Federal do Espírito Santo.

Vargas, R.E.V., Munaro, C.J., Ciarelli, P.M., Medeiros, A.G., do Amaral, B.G., Barrionuevo, D.C., de Araújo, J.C.D., Ribeiro, J.L., and Magalhães, L.P. (2019). A realistic and public dataset with rare undesirable real events in oil wells. *Journal of Petroleum Science and Engineering*, 181, 106223.