UNIVERSIDADE FEDERAL FLUMINENSE

ESCOLA DE ENGENHARIA

DEPARTAMENTO DE ENGENHARIA DE PRODUÇÃO

RENATO LOPES ROSA

CLASSIFICAÇÃO DE EVENTOS INDESEJAVEIS NA PRODUÇÃO DE PETRÓLEO *OFFSHORE* COM APLICAÇÃO DE TÉCNICAS DE INTELIGÊNCIA ARTIFICIAL

Niterói, RJ

2020

RENATO LOPES ROSA

CLASSIFICAÇÃO DE EVENTOS INDESEJAVEIS NA PRODUÇÃO DE PETRÓLEO *OFFSHORE* COM APLICAÇÃO DE TÉCNICAS DE INTELIGÊNCIA ARTIFICIAL

Trabalho de Conclusão de Curso apresentado ao Corpo Docente do Departamento de Engenharia de Produção da Escola de Engenharia da Universidade Federal Fluminense, como requisito parcial à obtenção do título de Bacharel em Engenharia de Produção.

Orientador(a):

Prof. Dr. Gilson Brito Alves Lima

Niterói, RJ

2020

Ficha catalográfica automática - SDC/BEE
Gerada com informações fornecidas pelo autor

R788c Rosa, Renato Lopes
CLASSIFICAÇÃO DE EVENTOS INDESEJAVEIS NA PRODUÇÃO DE PETRÓLEOOFFSHORE COM APLICAÇÃO DE TÉCNICAS DE INTELIGÊNCIA ARTIFICIAL / Renato Lopes Rosa ; Gilson Brito Alves Lima, orientador. Niterói, 2020.
48 f. : il.

Trabalho de Conclusão de Curso (Graduação em Engenharia de Produção)-Universidade Federal Fluminense, Escola de Engenharia, Niterói, 2020.

1. Inteligência Artificial. 2. Algoritmos de classificação. 3. Produção offshore. 4. Produção intelectual. I. Lima, Gilson Brito Alves, orientador. II. Universidade Federal Fluminense. Escola de Engenharia. III. Título.

CDD -

Bibliotecário responsável: Sandra Lopes Coelho - CRB7/3389

RENATO LOPES ROSA

CLASSIFICAÇÃO DE EVENTOS INDESEJAVEIS NA PRODUÇÃO DE PETRÓLEO *OFFSHORE* COM APLICAÇÃO DE TÉCNICAS DE INTELIGÊNCIA ARTIFICIAL

Trabalho de Conclusão de Curso apresentado ao Corpo Docente do Departamento de Engenharia de Produção de Engenharia da Universidade Federal Fluminense, como requisito parcial à obtenção do título de Bacharel em Engenharia de Produção.

Aprovado em 10 de dezembro de 2020.

BANCA EXAMINADORA

_____

Prof. Dr. Gilson Brito Alves Lima - UFF

Orientador

_____

Profª. Drª. Nissia Carvalho Rosa Bergiante - UFF

_____

Profª. Drª. Fernanda Abizethe de Carvalho Duim - UFF

Niterói

2020

AGRADECIMENTOS

A meu pai, Carlos Renato, e a minha mãe, Sonia Cristina, pelo suporte e apoio ao longo dessa longa jornada sem eles, eu nada seria.

A meus amigos que trilharam essa comigo e a fizeram mais leve e divertida.

A meu orientador e mentor Gilson pelo apoio, aprendizado, exemplo e orientação nesse projeto, assim como nos projetos de iniciação cientifica e pesquisa.

À professora Nissia pela inspiração e amor a sua profissão.

Ao CENPES pelo suporte financeiro do projeto PD&I - ANP 21354-6 e o grande aprendizado decorrente do projeto.

Ao Projeto de Iniciação Científica (PIBIC - CNPq) “Formulações de Inteligência Artificial na Gestão Tecnológica” pelo conhecimento e oportunidade.

RESUMO

A indústria do petróleo é a principal matriz energética da sociedade atual, o aumento da população e de consumo de energia vêm gerando uma demanda constante de novas fontes de energia. Dentro deste cenário, a exploração de petróleo em alto mar veio como uma nova fonte de combustíveis fósseis. A utilização de inteligência artificial vem sendo adotada amplamente pela indústria em geral para diminuir custos e otimizar processos produtivos. Neste trabalho, dados de alguns sensores e válvulas dentre os inúmeros presentes em uma operação de extração de óleo e gás offshore são utilizados. Nesse sentido, o objetivo do presente trabalho é desenvolver uma modelagem, utilizando dados públicos obtidos a partir de Vargas et al. 2019, para classificar os eventos indesejados na produção de petróleo offshore por meio da aplicação de técnicas de inteligência artificial com o auxílio das ferramentas: *Support Vector Machine* (SVM) e *Random Forest* (RF). Para tal foi utilizado como ferramenta a linguagem Python para aplicação dos algoritmos.

Palavras-Chave: Inteligência Artificial, algoritmos de classificação, Produção *offshore*.

ABSTRACT

The oil industry is the main energy matrix of today's society, the increase in population and energy consumption has been generating a constant demand for new sources of energy, oil exploration on the high seas has come as a new source of fossil fuels. The use of artificial intelligence has been widely adopted by the industry in general to reduce costs and optimize production processes. This work uses data from some sensors and valves are used among the countless ones present in an oil and gas extraction operation. In this sense, the objective of the present work is to develop a modeling, using public data obtained from Vargas et al. 2019, to classify undesired events in the production of offshore oil through the application of artificial intelligence techniques with the aid of the tools Support Vector Machine (SVM) and Random Forest (RF). For this, Python was used as a tool to apply the algorithms.

Keywords: Artificial Intelligence, Classification algorithms, Offshore production.

LISTA DE ILUSTRAÇÕES

Figura 1 Linha de produção offshore. 13
Figura 2 Participação Energética. 14
Figura 3 Hidrato na linha de produção. 16
Figura 4 Nó de Decisão. 21
Figura 5 Gráfico de Entropia. 22
Figura 6 Random Forest. 23
Figura 7 Hiper plano de classificação SVM. 24
Figura 9 Vetores de suporte SVM. 24
Figura 10 Representação dos metodos de classificação. 25
Figura 11 CRISP DM.. 26
Figura 12 Imagem da Árvore de natal. 31
Figura 13 Algoritmo de separação dados de classificação e dados de treino. 35
Figura 14 Criação do classificador RF. 35
Figura 15 Realizando Treinamento. 36
Figura 16 Realizando Pedição. 36
Figura 17 Metricas de Avaliação de Desempenho. 36
Figura 18 Matriz Confusão probema BSW. 39
Figura 19 Matriz Confusão probema Fechamento DHSV. 40
Figura 20 Matriz Confusão probema perda repentina de produtividade. 41
Figura 21 Matriz Confusão probema hidrato. 41
Figura 22 Matriz confusão para conjunto de problemas. 42

LISTA DE TABELA

Tabela 1 Tabela da base de dados. 29
Tabela 2 Variáveis presentes na base de dados 28
Tabela 3 Classes encontradas para cada poço no conjunto de dados. 30
Tabela 4 Analise de qualidade de dados dos arquivos. 32
Tabela 5 Conjunto de dados com valores faltentes. 33
Tabela 6 Conjunto de dados com as variáveis a serem utilizadas 34
Tabela 7 Conjunto de dados sem instâncias com dados faltososo 34
Tabela 8 Saida de dados Matrix de Confusão 37
Tabela 9 Resultados das metricas de avaliação. 38
Tabela 10 Resultado das métricas de avaliação para o probema BSW. 39
Tabela 11 Resultado das métricas de avaliação para o problema fechamento DHSV. 40
Tabela 12 Resultado das métricas de avaliação para o problema da perda repentina de produtividade 41
Tabela 13: Resultado das métricas de avaliação para o problema hidrato. 42
Tabela 14: Resultado das métricas de avaliação para conjunto de problemas. 43

LISTA DE SIGLAS

ABNT	Associação Brasileira de Normas Técnicas
SVM	Suport Vector Machine
RF	Random Forest
IA	Inteligência Artificial
ANP	Agência Nacional do Petróleo
NPT	Non-production Time
AM	Aprendizado de Máquina
DHSV	*Down Hole Safety Valve*
BSW	*Basic Sediment Water*

SUMÁRIO

1 O PROBLEMA 13

1.1 INTRODUÇÃO 13

1.2 SITUAÇÃO PROBLEMA 15

1.2.1 Hidrato 15

1.2.4 Perda repentina de produtividade 17

1.3 OBJETIVOS 18

1.3.1 Objetivo Geral 18

1.3.2 Objetivos específicos 18

1.4 QUESTÕES DA PESQUISA 18

1.5 DELIMITAÇÃO DO ESTUDO 18

1.6 IMPORTÂNCIA DO ESTUDO 19

1.7 ORGANIZAÇÃO DO ESTUDO 19

2 CAPÍTULO 20

2.1 TÉCNICAS DE INTELIGENCIA ARTIFICIAL 20

2.2 *MACHINE LEARN (APRENDIZADO DE MÁQUINA)* 20

2.3 ÁRVORE DE DECISÃO 21

2.4 *RANDOM FOREST* 23

2.4.1 Máquina de Vetor de Suporte ou *Support Vector Machine* (SVM) 23

2.5 *CROSS INDUSTRY STANDARD PROCESS FOR DATA MINING (CRISP-DM)* 25

3 MATERIAIS E MÉTODOS 28

3.1 ENTENDIMENTO DO NEGÓCIO 28

3.2 ENTENDIMENTO DOS DADOS 28

3.3 PREPARAÇÃO DOS DADOS 33

3.4 MODELAGEM 35

4 AVALIAÇÃO 36

5 ANÁLISE E DISCUSSÃO DE RESULTADOS 38

5.1 EXPERIMENTO 1 39

5.2 BSW 39

5.2.1 DHSV 40

5.2.2 Perda Repentina de Produtividade 40

5.2.3 Hidrato 41
5.2.4 Experimento 2 42
6 CONCLUSÕES..... 44
6.1 RESPOSTAS ÀS QUESTÕES DE PESQUISA 44
6.2 DESDOBRAMENTOS FUTUROS DA MODELAGEM DESENVOLVIDA 44

1 O PROBLEMA

1.1 INTRODUÇÃO

O desenvolvimento da sociedade e novas tecnologias acompanham um aumento contínuo de demanda de energia. Para suprir tal demanda, a produção de petróleo aumenta a cada dia. Para tornar esse aumento possível, diversas técnicas de produção e exploração foram e são desenvolvidas, assim como novas tecnologias. Com o domínio de fontes energéticas não renováveis, o aumento de produção é baseado em descoberta e exploração *offshore* (GOMES e BARATA, 2007).

A produção de petróleo *offshore* demanda um alto nível de conhecimento e competências industriais devido ao elevado nível de desafios tecnológicos. Operações em mar aberto e em altas profundidades apresentam problemas únicos desse tipo de exploração. A alta complexidade deste processo gera um aumento de custos que pode ser de milhares de dólares por dia de produção (BAKER, 1998) (THOMAS, 2001). A fim de exemplificar o sistema empregado neste tipo de exploração, a imagem a seguir representa a linha de produção *offshore.*

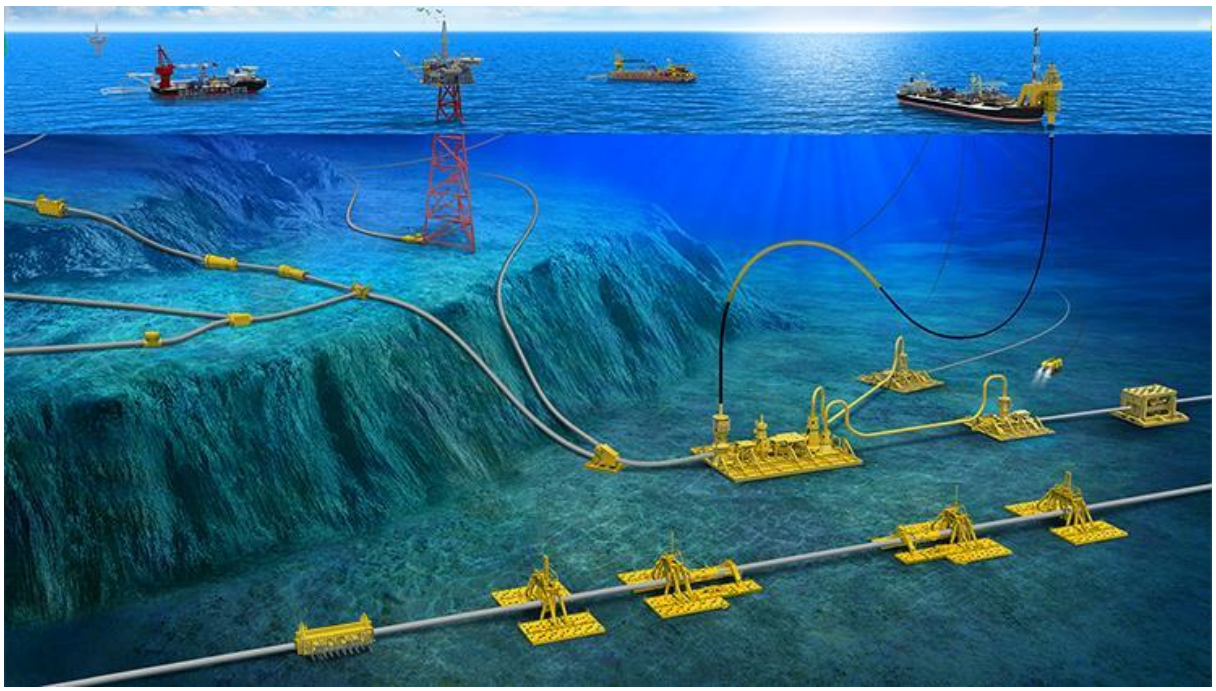

Figura 1*:* Linha de produção *offshore*. Fonte: Centro Brasileiro de Infraestrutura (2019).

Segundo o *Energy Information Administration* (2019) dos E.U.A, o petróleo e o gás

natural são as maiores fontes de energia mundial e suas projeções, como observadas na fig. 2, indicam que nos próximos 30 anos, apesar do aumento proporcional de energias renováveis, o petróleo e o gás natural irão se manter como a principal fonte de energia.

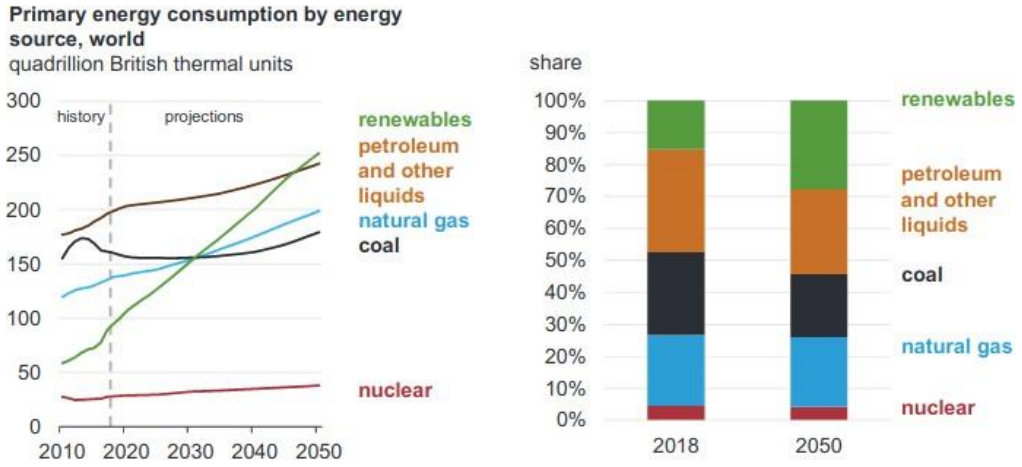

Figura 2: Participação Energética. Fonte: US. Energy Information Administration (2019).

Em 1938, o governo brasileiro criou o Conselho Nacional do Petróleo a fim de discutir o uso e a exploração do petróleo em terras brasileiras. Em 1953, a criação da estatal Petróleo Brasileiro S.A, Petrobras, foi anunciada. Como as jazidas de petróleo brasileira estão localizadas em águas profundas, a partir de 1968, a empresa começou a desenvolver um projeto para explorar esse petróleo (SOUSA, 2020).

O Brasil, atualmente, exploração petróleo *onshore* e *offshore*. Contudo, as principais jazidas estão localizadas em alto mar, os campos marítimos, os quais representam uma proporção de 95,9% da produção de petróleo e 82,1% da produção do gás natural (ANP, 2019). Salienta-se que nessa área de reserva petrolífera marítima localiza-se o pré-sal, o qual corresponde com uma parcela do petróleo brasileiro explorado em águas profundas.

Este sistema de produção, segundo Vargas *et al*. (2019), pode apresentar inúmeros problemas, a saber: aumento de sedimentos e água na linha de produção; fechamento da válvula de segurança; instabilidade de fluxo; formação de hidrato; entre outros.

Dentre esses problemas, a formação de hidrato é considerada um dos maiores na indústria do petróleo, pois ocorre com mais frequência em ambientes de alta pressão e baixa temperatura, características essas da produção de óleo e gás no pré-sal.

As ocorrências indesejadas desses eventos têm potencial para diminuir a produção durante dias e até semanas, o que causa espaços de tempo não produtivo (NPT[1]), os quais contribuem, consideravelmente, para a ocorrência de paradas da linha de produção.

1.2 SITUAÇÃO PROBLEMA

Dentro desse sistema produtivo, a ocorrência de eventos indesejados pode ocasionar a perda de produtividade ou até mesmo avaria de equipamentos, tais eventos podem ser identificados como interrupções nos dutos de transporte do óleo ou variação em taxas de sedimentos, bem como o comportamento indevido da válvulas. Dentre estes problemas, encontra-se o hidrato.

1.2.1 Hidrato

O hidrato é uma estrutura cristalina, com dois ou mais componentes, originária da junção de água com hidrocarbonetos submetidos a alta pressão e baixa temperatura ou gases de baixo peso molecular. Este fenômeno ocorre dentro dos dutos de produção, onde o gás e água escoam, como mostrado na fig. 3. Devido à força de ligação das pontes de hidrogênio na molécula de água, um retículo cristalino origina-se, geralmente, no momento em que a água encapsula a molécula de gás. No escoamento do óleo, a formação de hidrato inicia-se quando a água entra em contato com o gás, neste momento uma película de hidrato é formada isolando a fase água da fase óleo. Nesse sentido, a aglomeração desses fenômenos pode formar um *plug* de hidrato (SANTOS, DANTAS, *et al.*, 2020).

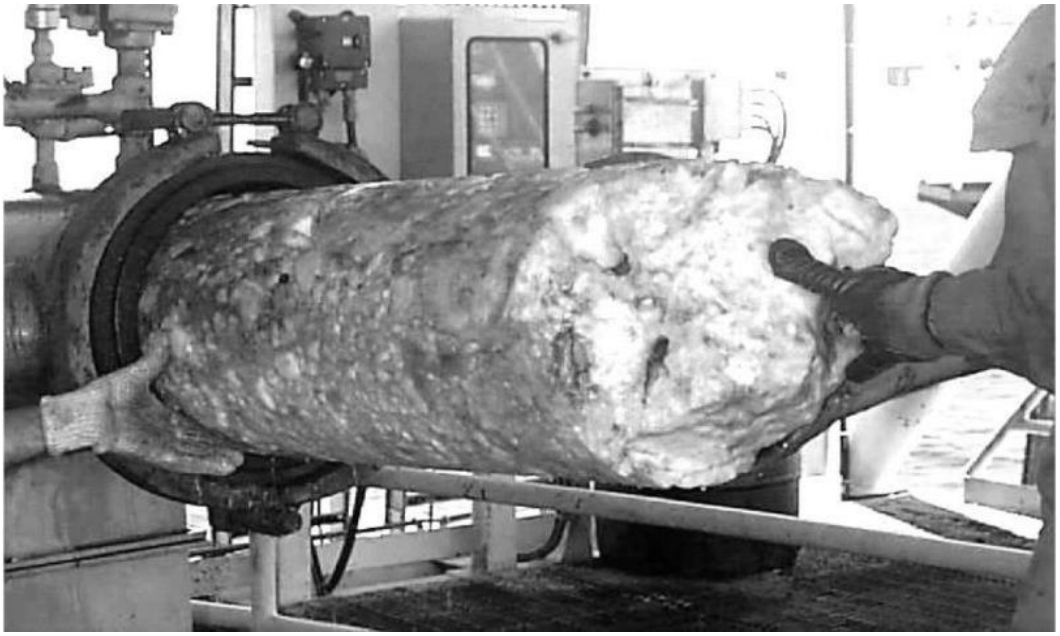

Figura 3: Hidrato na linha de produção. Fonte: Podorozhnikov, Shabarov, *et al.* (2016).

1.2.2 Fechamento da *Down Hole Safety Valve* sem indicação.

O *Down Hole Safety Valve* (DHSV), também referenciado como DSV, é uma válvula de segurança instalada na turbina de produção do poço de petróleo. Seu objetivo é garantir o fechamento do poço em caso de desconexão entre a unidade de produção e o poço de petróleo, de uma emergência ou evento de falha grave no equipamento de superfície. Portanto, ela é configurada de forma a interromper o fluxo na falha do sistema (SCHLUMBERGE, 2018), (STANDARDS NORWAY, 2013).

Eventualmente, essa função de fechamento de falha ocorre de uma maneira *spurius*, isto é, sem nenhuma indicação na superfície. A identificação automática desse fechamento em um tempo hábil pode possibilitar a reabertura por meio de medidas corretivas, assim evitando perdas e custos adicionais (VARGAS et al,2019).

1.2.3 Taxa *Basic Sediment and water* indesejada.

Outro elemento de segurança é a taxa *Basic Sediment and water* (BSW), o qual é definida como uma razão entre fluxo de sedimentos e água ambos medidos em condições normais de pressão e temperatura (NTP) (ANDREOLLI, 2016), (ABASS AND BASS, 1988). Durante o ciclo de vida de um poço de petróleo, espera-se que a taxa BSW aumente devido à

produção de água do aquífero ou da injeção artificial, a qual eleva o óleo e mantém a taxa de produção. Uma variação repentina do BSW, no entanto, pode ocasionar uma série de problemas relacionados ao fluxo, ao decréscimo na taxa de produção, ao levantamento de óleo, à incrustação, ao processamento na planta, e à taxa de recuperação. A identificação automática desses eventos indesejados permite realizar ações na administração da produção ou na injeção artificial a fim de evitar esses tipos de problemas (VARGAS et al,2019).

1.2.4 Perda repentina de produtividade

A produtividade de poços de petróleo de fluxo natural depende de várias propriedades, tais como: pressão estática do reservatório, porcentagem de sedimentos e água, viscosidade do líquido produzido, diâmetro da linha de produção, entre outras (HAUSLER et al., 2015).

Dependendo da alteração dessas propriedades, a linha de produção apresenta perda de fluxo. Desse modo, a identificação automática dessa variação de fluxo pode permitir que o time de operação intervenha no poço com o intuito de manter a produtividade (VARGAS et al,2019)

A predição desses eventos anômalos, a partir de técnicas de inteligência artificial, tem sido uma busca constante pelas organizações. No entanto, a grande deficiência das organizações é a ausência de dados históricos para embasar a modelagem, bem como de profissionais especializados que dominem, simultaneamente, tanto as técnicas de IA quanto os processos produtivos em análise. A fim de predizer esses eventos anômalos, há uma necessidade de utilizar os dados históricos oriundos tanto das medições de válvulas quanto do conjunto de sensores.

As vantagens da utilização de técnicas de inteligência artificial podem ser identificadas na redução de custos operacionais, uma vez que ao identificar problemas e mitigar seus danos, resultará numa maior eficiência, tendo em vista a diminuição do tempo improdutivo de uma planta de produção. Além disso, a automação de tarefas, ao atribuir ao algoritmo a realização de algum trabalho que seria realizado por pessoas, aumentará a agilidade na resolução de problemas, como a identificação precoce da origem do problema.

Neste sentido, a aplicabilidade do estudo está associada ao preenchimento da lacuna existente - tanto pela incipiência de trabalhos no contexto da Engenharia de Produção, quanto da escassez na literatura especializada - de aplicações correlatas de técnicas de inteligência artificial para interpretação dos fenômenos de classificação e predição.

1.3 OBJETIVOS

1.3.1 Objetivo Geral

O objetivo do presente trabalho é desenvolver uma modelagem para classificar eventos indesejados na produção de petróleo *offshore* por meio da aplicação de técnicas de inteligência artificial.

1.3.2 Objetivos específicos

- Analisar algoritmos de inteligência artificial aplicáveis ao problema de estudo.
- Identificação dos dados necessários para a aplicação das técnicas de inteligência artificial.
- Tratar os dados e modelar o problema.
- Analisar os indicadores de qualidade.

1.4 QUESTÕES DA PESQUISA

- Qual ferramenta utilizar para classificar o problema anômalo em estudo?
- Quais as variáveis fenomenológicas aplicáveis?
- Como utilizar os resultados da pesquisa?

1.5 DELIMITAÇÃO DO ESTUDO

Diversos fenômenos anômalos ocorrem no processo de produção de óleo e gás (VARGAS *et al*., 2019). Dentre os fenômenos mais recorrente de produção podem ser citados a instabilidade de fluxo, a perda rápida de produtividade, *scaling* no PCK, a restrição do PCK hidrato na linha de produção.

Dado esse cenário de complexos fenômenos de distintas origens e variáveis, o presente estudo se delimitará a classificação de eventos anômalos na produção *offshore*, no contexto da alteração do estado normal para anormal.

1.6 IMPORTÂNCIA DO ESTUDO

O desenvolvimento e a utilização de ferramentas de controle auxiliam a empresa a melhor sua produtividade e, consequentemente, sua competitividade no mercado. Com os atuais desenvolvimentos tecnológicos na área de análise de dados, o tema se torna, extremamente, relevante para o engenheiro de produção.

1.7 ORGANIZAÇÃO DO ESTUDO

O projeto final de curso se encontra dividido em cinco capítulos, a saber: o primeiro capítulo faz a contextualização do mercado de petróleo para a matriz energética mundial, em seguida descreve-se o objetivo do projeto, delimitando-o e mostrando a importância no contexto de engenharia de produção; o segundo capítulo apresenta a base de conhecimento teórico necessário para o entendimento, o desenvolvimento e a análise do estudo; o terceiro capítulo apresenta a aplicação da metodologia proposta para o desenvolvimento do trabalho.

2 CAPÍTULO

2.1 TÉCNICAS DE INTELIGENCIA ARTIFICIAL

O termo técnicas de inteligência artificial remonta a 1943 cunhado por McCulloch e Pitt, eles apresentaram um trabalho pioneiro, Inteligência Artificial (IA). Esse artigo demonstra que as máquinas artificiais possuem inteligência similar à de humanos. Com o objetivo de capacitar os computadores para executar funções desempenhada pelo ser humano por meio do conhecimento e raciocínio, o requisito fundamental para a existência de sistemas inteligentes é a assimilação de conhecimento (MITCHELL, 1997) (REZENDE, 2003). Turing, em seu artigo "Computação de Máquina e Inteligência" (1950), apresenta o teste de Turing, o qual preconiza aprendizado de máquina, algoritmos genéticos e reforço de aprendizagem. Além disso, apresenta uma definição operacional da inteligência artificial (RUSSELL, S. J.; NORVIG, P., 2010).

2.2 *MACHINE LEARN (APRENDIZADO DE MÁQUINA)*

O aprendizado de máquina (AM) consiste de programas de computador que por meio da experiência melhoram seu desempenho em determinada tarefa (MICHALSKI, CARBONELL e MITCHELL, 2013). Segundo Nasrabadi (2007), a AM visa a resolver problemas por meio de modelados de variáveis, os quais utilizam funções ou conjunto de funções. A fim de explicar as amostras do problema, as etapas anteriores são utilizadas por este conjunto de algoritmos.

A capacidade de melhorar o desempenho de alguma tarefa através da experiência também é a definição literal de Aprendizado de máquina (FACELI, LORENA, *et al.*, 2011)

Existe diferentes tipos de variáveis para representar as características do problema, estas podem ser contínuas, binárias ou categóricas. Existe duas formas de aprendizado de máquina: o supervisionado, as saídas corretas são informadas ao algoritmo, e o não supervisionado, as saídas corretas não são informadas (KOTSIANTIS, ZAHARAKIS e PINTELAS, 2007).

Problemas de regressão e classificação podem ser resolvidos pelo aprendizado supervisionado. Os problemas de classificação, por exemplo, seriam categorizar frutas de

acordo com as características delas, como cor ou rugosidade. Por sua vez, os problemas de regressão seriam a predição de um valor numérico contínuo desconhecido em cima de um conjunto ordenado de valores (FACELI, LORENA, *et al.*, 2011). Salienta-se que o foco deste trabalho consiste em problemas de classificação.

2.3 ÁRVORE DE DECISÃO

As árvores de decisão trabalham por meio da divisão de problemas complexos em problemas simples, o qual utiliza a estratégia de dividir para conquistar (FACELI, LORENA, *et al.*, 2011).

A árvore de decisão é caracterizada por apresentar etapas e processos. O problema inicial é denominado raiz. Os nós de decisão têm a função de realizar o teste de um dos atributos, o que, por sua vez, gera ramos descendentes que correspondem a um possível valor do atributo associado a este nó. Além disso, há também as folhas que estão associadas a uma classe. Cada caminho percorrido da folha à raiz representa uma regra de classificação (FACELI, LORENA, *et al.*, 2011)

A divisão do espaço de predição ocorre em divisões sequências binárias nos nós, os que não são divididos são chamados nós finais, e é a última parte do espaço de predição que consiste no caminho da raiz ao nó final. Cada nó não terminal se divide em dois nós descendentes de acordo com o valor da variável de predição, o ponto de separação entre esses nós pode ser o valor de uma variável, sendo maior que determinado valor. Caso afirmativo, ele vai para um nó e caso negativo, para outro (BREIMAN, 2001).

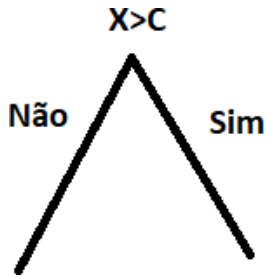

Figura 4: Nó de Decisão.

Dentre as vantagens da utilização desse método encontra-se o fato dele ser não-

paramétrico, isto é, não assume distribuição particular para os dados e pode construir modelo para qualquer função com um número suficiente de exemplos de treino, a quantidade de pontos para classificar um problema depende da complexidade do problema, a decomposição sucessiva do problema em decisões elementares garante uma grande interpretabilidade do problema, muito eficiente para construir modelos (FACELI, *et al*, 2011).

Entropia é a medida de aleatoriedade na informação a ser processada. Quanto maior a entropia mais difícil classificar a informação. O jogar de uma moeda pode ser um exemplo de ação que confirma sua aleatoriedade.

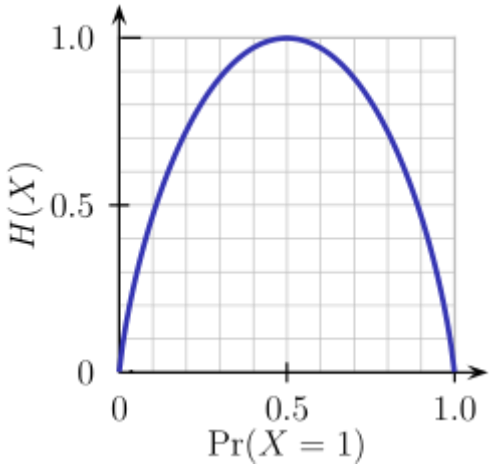

Figura 5: Gráfico de Entropia. Fonte: Freitas (2004).

Na figura 5 acima, a entropia H(x) é zero, quando a probabilidade é zero ou um, e apresenta seu valor máximo no 0,5, o qual demonstra sua aleatoriedade, podendo ser cara ou coroa com a mesma probabilidade.

$$\boldsymbol{Entropia(S)} = \sum \boldsymbol{p_i} \log_2 \boldsymbol{p_i}$$

Equação 1: Cálculo de Entropia (FREITAS, 2004).

O ganho de informação ou GI é uma propriedade estatística que mede quão bem determinado atributo separa os exemplos de treino de acordo com sua classificação. Matematicamente a GI é representada pela equação 2 sendo:

$$GI\ (T, X) = Entropia(T) - Entropia\ (T, X)$$

Equação 2: Cálculo de Ganho (FREITAS, 2004).

2.4 *RANDOM FOREST*

Random Forest consiste em uma combinação de inúmeras árvores aleatórias para criar um único modelo, desta forma cada árvore de decisão criada será comparada com outra para encontrar melhor representação do problema. Por conseguinte, resulta em uma melhor classificação (KOEHRSEN, 2020).

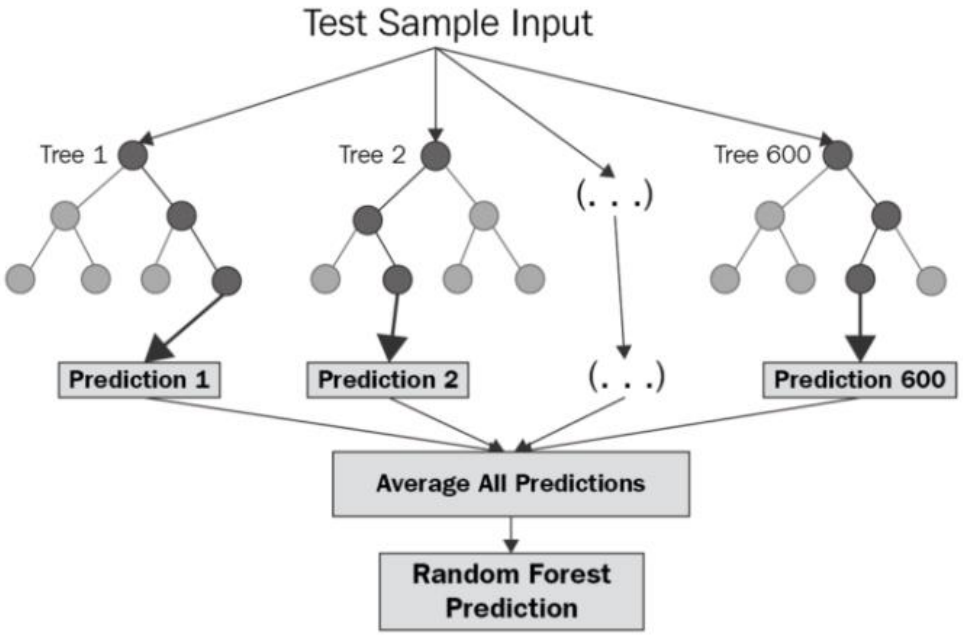

Figura 6: *Random Forest*. Fonte: Chakure (2019).

2.4.1 Máquina de Vetor de Suporte ou Support Vector Machine (SVM)

O *Support Vector Machine* (SVM) é um algoritmo supervisionado que pode ser usado para problemas de classificação e regressão. A ideologia do método é baseada na ideia de se encontrar um hiperplano que melhor separa as classes em domínios diferentes. Um exemplo de utilização é a classificação de e-mails, em que um hiperplano vai separar duas classes.

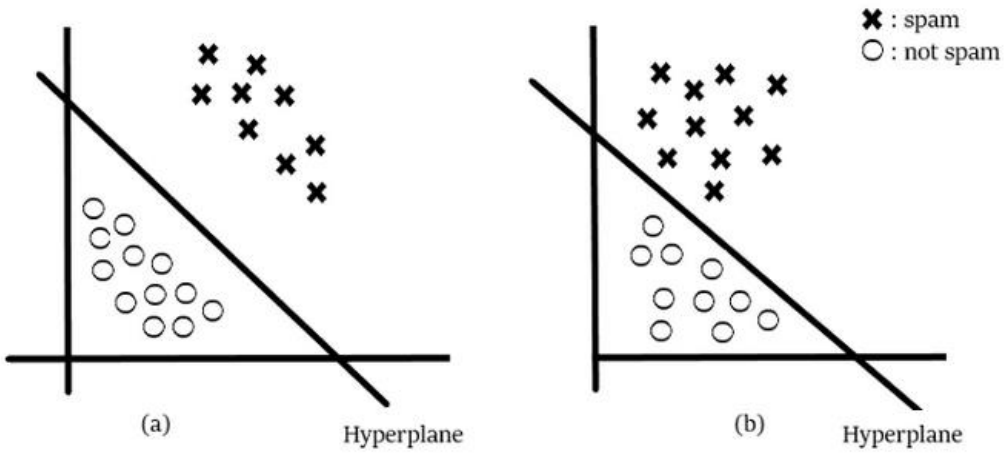

Figura 7: Hiper plano de classificação SVM. Fonte: Yadav (2020).

A melhor opção de hiperplano é identificada por meio da medição da distância entre os pontos de suporte do vector. Essa distância entre os pontos e o hiperplano é chamado de margem. Os pontos de suporte são críticos para determinar qual hiperplano é o ideal, pois a alteração dos pontos de vector resulta na alteração da posição do hiperplano.

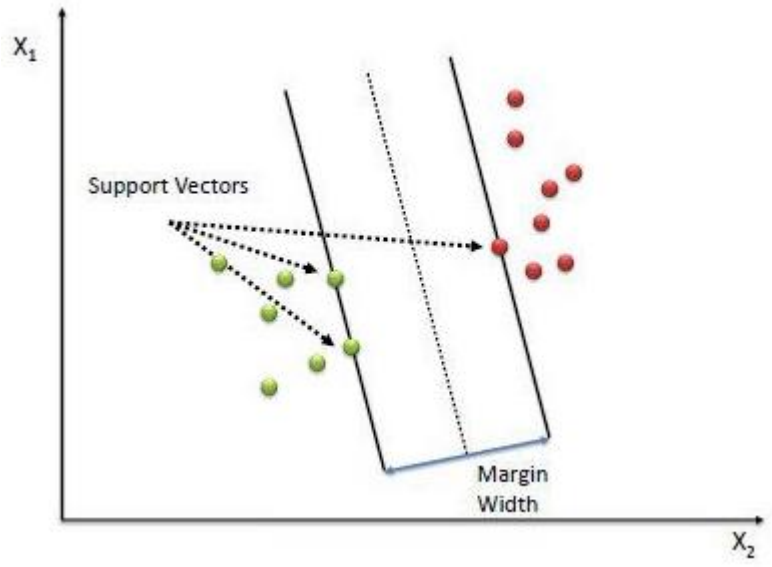

Figura 8: Vetores de suporte SVM (YADAV, 2020).

O hiperplano consiste em uma função que separa duas classes em uma classificação em duas dimensões a função seria uma reta, por sua vez, uma classificação em três dimensões seria um plano e receberia o nome de plano de similaridade. Além disso, a função que classifica

pontos em dimensões superiores é denominada hiperplanos. Nesse sentido, a função de classificação utilizada no algoritmo pode ser baseada em 3 núcleos, linear, polinomial ou Gaussiano (YADAV, 2020).

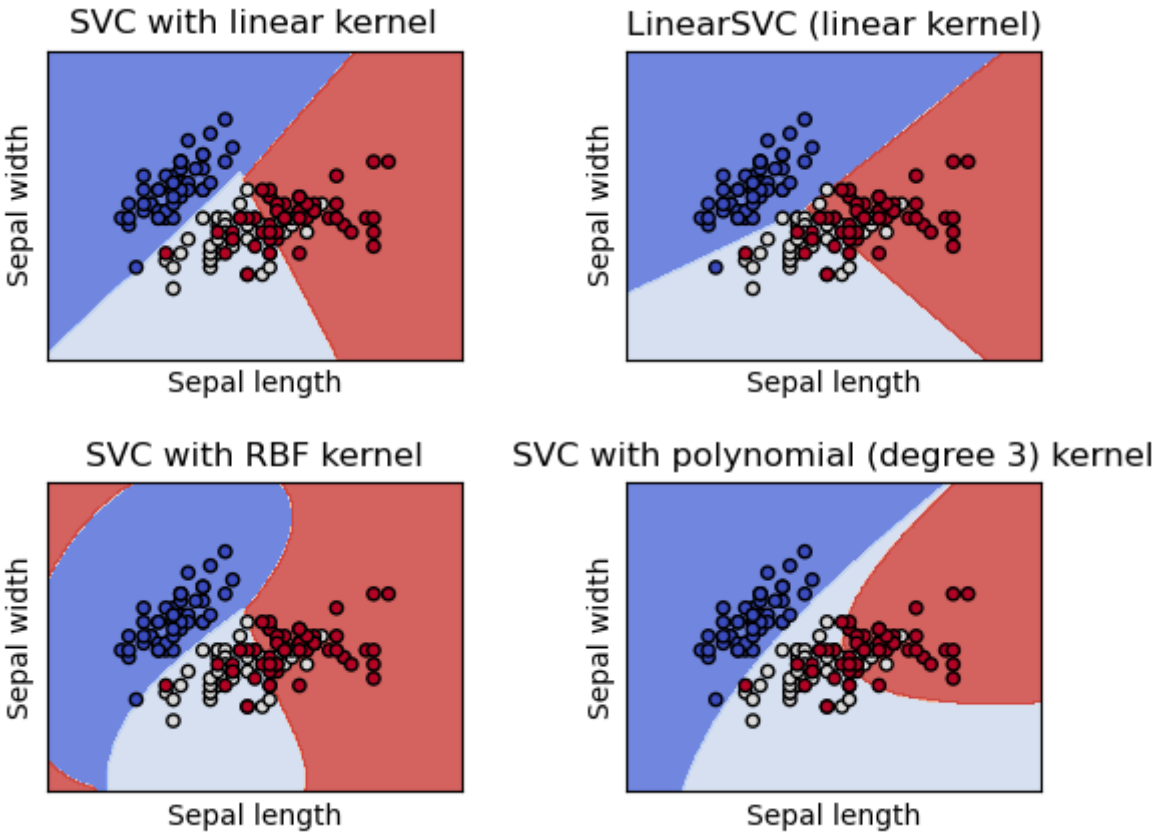

Figura 9: Representação dos métodos de classificação (PEDREGOSA, VAROQUAUX, *et al.*, 2011).

A diferença entre os métodos é a função utilizada para definir os intervalos de classificação. No atual estudo, será utilizado a metodologia Gaussiana (RBF).

2.5 *CROSS INDUSTRY STANDARD PROCESS FOR DATA MINING (CRISP-DM)*

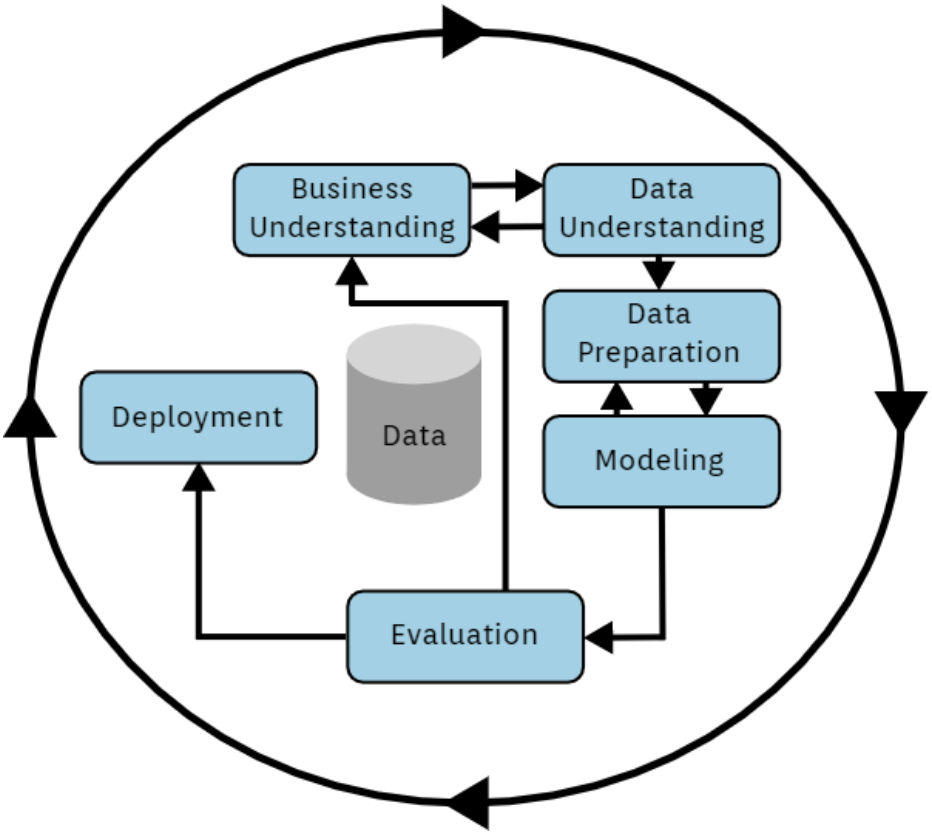

Figura 10: CRISP DM. Fonte: Dylan (2016).

O *Cross Industry Standard Process for Data Mining* (CRISP DM) é uma metodologia dividida em etapas para auxiliar e organizar o desenvolvimento de um projeto de mineração de dados, como demonstrado na fig. 10. Este método garante uma visão geral do ciclo de vida do projeto, contendo fases, tarefas e saída de dados do mesmo.

O ciclo de vida do projeto é separado em seis etapas, a sequência dessas etapas não é rigorosa, dependendo do resultado de cada etapa a próxima entrará em execução. Tais etapas dependem do entendimento do negócio, a fase inicial do projeto tem como foco o entendimento dos objetivos e das necessidades, convertendo o conhecimento em um projeto preliminar para alcançar seus objetivos.

A segunda etapa consiste em entender os dados, esta etapa se baseia em coletar e trabalhar os dados para maior familiaridade para identificar problemas de qualidade, *insights* ou detectar informações escondidas. Esta etapa está intimamente ligada com a etapa anterior, pois a formulação do problema de mineração de dados está relacionada com o entendimento do negócio.

A terceira etapa do processo é a preparação dos dados, esta contempla todas as atividades necessárias para construir o conjunto de dados que será utilizado pela ferramenta de

modelagem, ocorrem processos como limpeza e transformação dos dados.

A quarta etapa é a modelagem, nesta fase ocorre a seleção e aplicação e calibração de técnicas de modelagem de dados.

A penúltima etapa é a avaliação do modelo que realiza a avaliação do modelo e revisão de etapas para garantir que os objetivos do negócio foram atingidos e se verifica se nenhum problema relacionado foi esquecido.

A última etapa desta metodologia consiste em adequar a saída de dados para melhor entendimento do público alvo, podendo ser um simples relatório ou um processo de mineração de dados que se repete.

3 MATERIAIS E MÉTODOS

A metodologia CRISP é utilizada como um ponto de partida para a estruturação teórica da mineração de dados, cada fase desta metodologia representada pela imagem apresenta um propósito diferente (DEBUSE, 2007).

3.1 ENTENDIMENTO DO NEGÓCIO

Esta etapa é uma das mais fundamentais no processo de mineração de dados estando muito ligada ao sucesso do projeto (DEBUSE, 2007).

No contexto industrial tem ocorrido uma demanda de aumento da produtividade, qualidade e eficiência energética, a detecção e classificação de eventos indesejados é relevante para atividades exercidas ou monitoradas por humanos. A tarefa de responder a eventos anormais envolve a detecção prematura destes eventos, o diagnóstico da origem do problema e a intervenção para retornar esse evento a condição normal. O poço de petróleo corresponde a um conjunto de sensores e componentes mecânico, pneumáticos e hidráulicos, tais componentes e válvulas são monitorados pela operadora do poço (VARGAS, MUNARO, *et al.*, 2019)

3.2 ENTENDIMENTO DOS DADOS

Esta etapa consistiu na visualização e identificação das variáveis coletadas pela Petrobras, os dados foram disponibilizados em forma de *Comma-Separated Values* (CSV) e foram disponibilizados pelo (VARGAS, MUNARO, *et al.*, 2019).

O arquivo utilizado apresenta uma classificação de estados, esses dados estão classificados em três grupos, a saber, normal, transiente e falha. Esses grupos apresentam os respectivos valores dos sensores, como podemos identificar na tabela 1, a qual o 0.0 na coluna *class* representa o estado normal e o 101 representa o transiente.

	timestamp	P-PDG	P-TPT	T-TPT	P-MON-CKP	T-JUS-CKP	P-JUS-CKGL	T-JUS-CKGL	QGL	class
0	2018-06-18 06:02:45.000000	-1.180116e+42	20787460.0	117.9124	10085640.0	70.86996	4096358.0	NaN	0.0	0.0
1	2018-06-18 06:02:46.000000	-1.180116e+42	20787460.0	117.9123	10085720.0	70.86991	4096360.0	NaN	0.0	0.0
2	2018-06-18 06:02:47.000000	-1.180116e+42	20787460.0	117.9122	10085810.0	70.86987	4096363.0	NaN	0.0	0.0
3	2018-06-18 06:02:48.000000	-1.180116e+42	20787460.0	117.9120	10085930.0	70.86981	4096365.0	NaN	0.0	0.0
4	2018-06-18 06:02:49.000000	-1.180116e+42	20787460.0	117.9119	10086050.0	70.86977	4096367.0	NaN	0.0	0.0
...	...	...	...	...	...	...	...	...	...	...
16960	2018-06-18 10:45:25.000000	-1.180116e+42	20619450.0	118.2390	9992133.0	71.26657	4108511.0	NaN	0.0	101.0
16961	2018-06-18 10:45:26.000000	-1.180116e+42	20619480.0	118.2390	9992676.0	71.26487	4108512.0	NaN	0.0	101.0
16962	2018-06-18 10:45:27.000000	-1.180116e+42	20619520.0	118.2390	9993218.0	71.26317	4108512.0	NaN	0.0	101.0
16963	2018-06-18 10:45:28.000000	-1.180116e+42	20619550.0	118.2390	9993761.0	71.26147	4108512.0	NaN	0.0	101.0
16964	2018-06-18 10:45:29.000000	-1.180116e+42	20619590.0	118.2390	9994304.0	71.25977	4108513.0	NaN	0.0	101.0

Tabela 1:Tabela da base de dados

Através da análise do arquivo base utilizado para essa pesquisa foram identificadas 8 variáveis P-PDG, P-TPT, T-TPT, P-MON-CKP, T-JUS-CKP, P-JUS-CKP, T-JUS-CKGL, QGL o significado de cada variável pode ser identificado na tabela 2.

Sigla	**Nome**	**Especificação**
PDG	*Permanent Downhole Gauge*	Sensor de pressão localizado no interior do poço
P-TPT	*Pressure at the Temperature and Pressure Transducer*	Sensor de pressão localizado na árvore de natal
T-TPT	*Temperature as the Temperature and Pressure Transducer*	Sensor de pressão localizado na árvore de natal
P-MON-CKP	*Pressure upstream the choke and production valve*	Sensor de pressão à montante da válvula choque de produção.
T-JUS-CKP	*Temperature downstream the choke production valve*	Sensor de temperatura à jusante da válvula choque de produção.
P-JUS-CKGL	*Pressure downstream choke gas lift valve*	Sensor de pressão à jusante da válvula choque de gás lift.
T-JUS-CKGL	*Temperature downstream choke gas lift valve*	Sensor de temperatura à jusante da válvula choque de gás lift.
QGL	*Gas lift flow*	Vazão de gás lift

Tabela 2:Variáveis presentes na base de dados (SANTOS, 2020, p. 56).

Vargas *et. al.* (2019) apresentou seis problemas, os quais estão identificado na tabela

3, dentro de cada problema, as classes apresentadas pelo mesmo, no qual o zero é sempre a condição normal de produção, o algarismo com três casas representa o transiente e o número inteiro de um algarismo representa as falhas.

	Normal (0)	**BSW (1)**	**DHSV (2)**	**LOSS (5)**	**RESTRI (6)**	**SCALING(7)**	**HYDRATE (8)**
Poço 1		0,101,1				0,107,7	
Poço 2		0,101,1	0,102,2		0,106,6		
Poço 3			0,102,2				
Poço 4					0,106,6		
Poço 5							
Poço 6		0,101,1				0,107	
Poço 7							
Poço 8							
Poço 9			0,102,2				
Poço 10			0,102,2				
Poço 11			0,102,2				
Poço 12			0,102				
Poço 13			0,102,2				
Poço 14							
Poço 15				0,105,5			
Poço 16				0,105,5			
Poço 17				0,105			
Poço18						107	
Poço 19							0,108,8
Poço 20							0,108,8
Poço 21							0,108,8

Tabela 3: Classes encontradas para cada poço no conjunto de dados.

O presente estudo utilizou somente os sensores presentes na arvore de natal, o posicionamento dos mesmos é demonstrado na figura 12. A arvore de natal se posiciona no leito marinho, desta forma seus sensores são identificados como sensores de fundo.

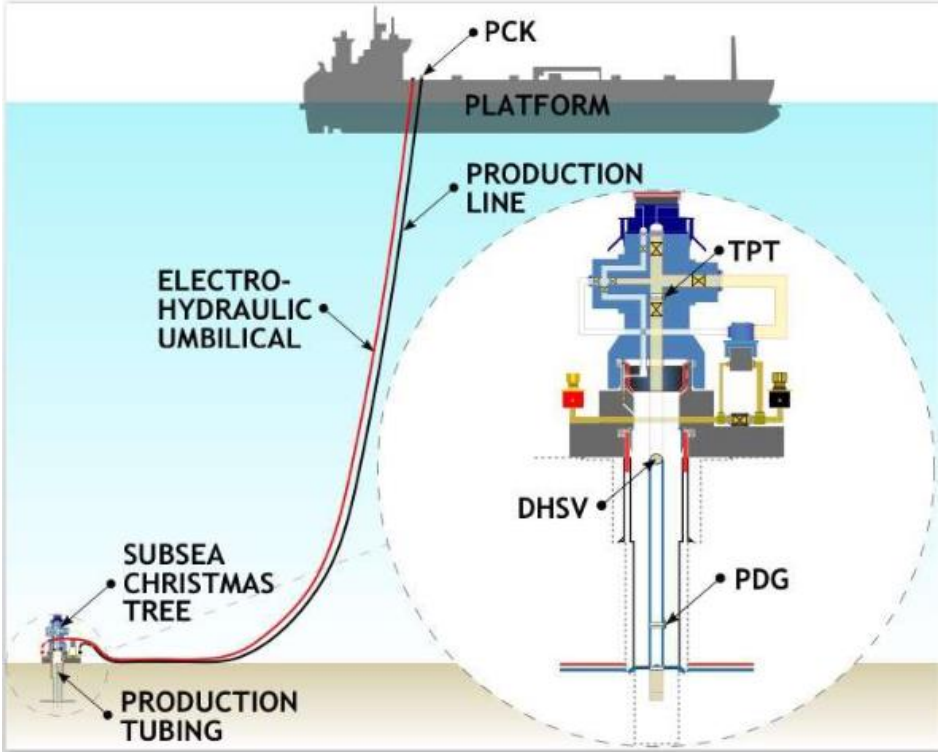

Figura 11: Imagem da Árvore de natal (VARGAS et al,2019).

Analisando os sensores presentes na árvore de natal foi avaliado a qualidade dos dados para os arquivos utilizados, para desta forma identificar os arquivos que apresentem leituras para os 3 sensores selecionados.

Os dados para cada poço foram considerados de boa qualidade caso apresentem as 3 classes para o problema em questão, na tabela 4 foi demonstrada a avaliação referente a cada poço. Foi considerado de baixa qualidade também os poços que apresentaram dados contendo as três classes porem para problemas que não serão abordados no presente estudo.

	PDG	**P-TPT**	**T-TPT**	**Quality**
Poço 1		X	X	OK
Poço 2		X	X	OK
Poço 3	X	X	X	OK
Poço 4		X	X	OK
Poço 5	-	-	-	Relacionado com problema 3 e 4
Poço 6	X	X	X	Baixa qualidade dos dados relacionados anomalia 7, não tem classe do problema.
Poço 7	-	-	-	Relacionado com problema 3 e 4

Poço 8	-	-	-	Apresenta somente classe normal
Poço 9		X	X	OK
Poço 10	X	X	X	OK
Poço 11	X	X	X	OK
Poço 12				*Dropped*
Poço 13		X	X	OK
Poço 14	-	-	-	Relacionado com problema 3 e 4
Poço 15				Baixa qualidade dos dados relacionados a anomalia 5, não tem classe do problema.
Poço 16	X	X	X	OK
Poço 17				Faltando dados
Poço18				Faltando dados
Poço 19				Faltando dados
Poço 20	X	X	X	OK
Poço 21	X	X	X	OK

Tabela 4: Analise de qualidade de dados dos arquivos. Fonte: o autor (2020).

Com a análise da qualidade dos dados foram identificados sete poços que apresentam as variáveis que serão utilizadas, 3,6,10,11,16,20,21, dentre esses poços foram identificados quatro problemas a serem abordados, estes são: a) hidrato; b) bsw; d) dhsv; e e) loss.

3.3 PREPARAÇÃO DOS DADOS

A qualidade dos dados afeta, diretamente, o processo de mineração de dados e tem impacto significativo no sucesso do projeto (DEBUSE, 2007).

Com o intuito de melhorar a qualidade dos dados algumas medidas foram tomadas como:

- A biblioteca pandas da linguagem *python foi* utilizada para fazer o tratamento.
- A normalização dos dados em cada uma das variáveis baseada no valor médio da classe normal, por sua vez as classes transientes e problema variam em torno do valor da classe normal, desta forma é possível analisar poços de diferentes intervalos de tempo, uma vez que o normal vai estar sempre variando perto de 1.
- A limpeza dos dados devido à falta de informação de alguns sensores ou em determinados períodos de tempo com o intuito de adequar os dados para a aplicação do RF e SVM.
- A união dos das classificações normal e anormal um uma única classe anormal.

A tabela 5 mostra os dados normalizados para o problema de hidrato, o tratamento de dados demonstrado foi o mesmo para todos os problemas descritos anteriormente.

	timestamp	P-PDG	P-TPT	T-TPT	P-MON-CKP	T-JUS-CKP	P-JUS-CKGL	T-JUS-CKGL	QGL	class
0	2017-03-01 18:23:17	NaN	NaN	NaN	1.000427	NaN	0.993199	NaN	1.177426e+00	Normal
1	2017-03-01 18:23:18	NaN	NaN	NaN	1.000426	NaN	0.993198	NaN	1.174636e+00	Normal
2	2017-03-01 18:23:19	NaN	NaN	NaN	1.000426	NaN	0.993197	NaN	1.171846e+00	Normal
3	2017-03-01 18:23:20	NaN	NaN	NaN	1.000425	NaN	0.993196	NaN	1.169057e+00	Normal
4	2017-03-01 18:23:21	NaN	NaN	NaN	1.000425	NaN	0.993195	NaN	1.166267e+00	Normal
...	...	...	...	...	...	...	...	...	...	...
90717	2017-05-09 08:18:36	1.006160	2.222192	0.238146	2.371027	NaN	1.130153	NaN	1.947591e-08	Abnormal
90718	2017-05-09 08:18:37	1.006159	2.221467	0.238150	2.371028	NaN	1.130113	NaN	1.947522e-08	Abnormal
90719	2017-05-09 08:18:38	1.006158	2.220744	0.238154	2.371029	NaN	1.130073	NaN	1.947453e-08	Abnormal
90720	2017-05-09 08:18:39	1.006158	2.220021	0.238159	2.371031	NaN	1.130032	NaN	1.947384e-08	Abnormal
90721	2017-05-09 08:18:40	1.006158	2.218964	0.238153	2.371031	NaN	1.129991	NaN	1.947314e-08	Abnormal

Tabela 5: Conjunto de dados com valores faltantes.

Após a seleção das três variáveis presentes nos arquivos que fazem parte do

conjunto da árvore de natal as demais variáveis tirando o P-PDG, P-TPT e T-TPT, foram excluídas do conjunto de dados, identificados na tabela 6.

	timestamp	P-PDG	P-TPT	T-TPT	class
0	2017-03-01 18:23:17	NaN	NaN	NaN	Normal
1	2017-03-01 18:23:18	NaN	NaN	NaN	Normal
2	2017-03-01 18:23:19	NaN	NaN	NaN	Normal
3	2017-03-01 18:23:20	NaN	NaN	NaN	Normal
4	2017-03-01 18:23:21	NaN	NaN	NaN	Normal
...	...	...	...	...	...
90717	2017-05-09 08:18:36	1.006160	2.222192	0.238146	Abnormal
90718	2017-05-09 08:18:37	1.006159	2.221467	0.238150	Abnormal
90719	2017-05-09 08:18:38	1.006158	2.220744	0.238154	Abnormal
90720	2017-05-09 08:18:39	1.006158	2.220021	0.238159	Abnormal
90721	2017-05-09 08:18:40	1.006158	2.218964	0.238153	Abnormal

Tabela 6: Conjunto de dados com as variáveis a serem utilizadas.

A partir da análise das variáveis selecionadas para realizar o presente estudo, foi identificado a falta de valores em alguns intervalos de tempo, a falta de informação para essa variável pode ocasionar um erro de treinamento da ferramenta para classificar os fenômenos, desta forma estas linhas de informação que apresentam valores faltantes serão desconsideradas durante o treinamento e classificação do algoritmo e conjunto de dados a ser utilizado esta visível na tabela 7.

timestamp	P-PDG	P-TPT	T-TPT	class
2012-04-10 19:23:26	0.999205	1.010811	0.963367	Normal
2012-04-10 19:23:27	0.999214	1.010596	0.963793	Normal
2012-04-10 19:23:28	0.999223	1.010381	0.964219	Normal
2012-04-10 19:23:29	0.999233	1.010165	0.964645	Normal
2012-04-10 19:23:30	0.999242	1.009950	0.965071	Normal
...	...	...	...	...
2017-05-09 08:18:36	1.006160	2.222192	0.238146	Abnormal
2017-05-09 08:18:37	1.006159	2.221467	0.238150	Abnormal
2017-05-09 08:18:38	1.006158	2.220744	0.238154	Abnormal
2017-05-09 08:18:39	1.006158	2.220021	0.238159	Abnormal
2017-05-09 08:18:40	1.006158	2.218964	0.238153	Abnormal

Tabela 7: Conjunto de dados sem instancias com dados faltosos.

A tabela 7 mostra o conjunto de dados somente com dados completos, ou seja, apresenta valores para todas as variáveis em todo intervalo de tempo.

3.4 MODELAGEM

O presente trabalho vai ser dividido em dois blocos de experimento, o primeiro bloco avaliará a utilização do algoritmo para classificação de classes nos arquivos de cada poço de petróleo, o segundo experimento vai consistir em um arquivo unificado por problema em que todos os poços que apresentam o mesmo problema serão unificados em um único arquivo que será submetido ao treinamento e teste do classificador.

Nesse momento, foi utilizada a biblioteca s*klearn* do *python* para utilizar os algoritmos RF e SVM. Por meio dos algoritmos, os dados foram segregados, 70% do volume desses dados tratados foram usados para realizar o treinamento, e 30% dos dados restantes para o teste de validação. O treinamento foi realizado com a importação de um arquivo CSV no algoritmo, desta forma o pacote *sklearn* gera um algoritmo de classificação com a capacidade de classificar em classes.

O primeiro passo após a limpeza dos dados é separá-lo em dois conjuntos de dados, esse conjunto de dados é separado de forma aleatória, um para treinamento e outro para avaliação do algoritmo, correspondentes respectivamente a 70% e 30 % do conjunto de dados.

```
# Cria dois conjuntos de dados, treino e teste
# Separados aleatoriamente em 70 e 30% dos dados respectivamente
train, test = df[df['is_train']==True], df[df['is_train']==False]
```

Figura 12: Algoritmo de separação dados de classificação e dados de treino. Fonte: o autor (2020).

Com os dois conjuntos de dados separados, aplica-se os métodos de treinamento na parte respectiva, como podemos identificar na imagem a seguir o classificador RF é criado dentro da instância python. Para o SVM foi utilizado o mesmo processo com o código completo no apêndice.

```
# Cria um classificador Random Forest. clf significa classificador
clf = RandomForestClassifier(n_estimators=500,n_jobs=2, random_state=0)
```

Figura 13: Criação do classificador RF. Fonte: o autor (2020).

Com o classificador criado, o próximo passo é utilizar o conjunto de dados previamente separado e treinar o classificador com essa entrada de dados.

```
# Treina o classificador pela classe
classifier = pd.DataFrame(clf.fit(train[features], y))
print(clf.fit(train[features], y))
```

Figura 14:Realizando Treinamento Fonte: Autoria do autor.

Ao realizar o treinamento, o passo seguinte é testar o classificador utilizando o segundo segmento de dados que foi separado para essa finalidade.

```
# Aplica o Classificador aos dados de test
# Que nunca foram vistos pelo algoritmos
clf.predict(test[features])
```

Figura 15: Realizando Pedição. Fonte: Autoria do autor.

Quando o classificador gera a predição para os pontos imputados devemos aplicar as métricas de avaliação do algoritmo.

```
# Cria a matriz confusão
confusion_matrix = pd.crosstab(teste1, preds, colnames=['Predição'], rownames=['Atual'])

# Calcula as metricas de analise do Algoritmo
from sklearn.metrics import f1_score
print(f1_score(teste1, preds, average='weighted'))

classific = classification_report(teste1, preds, target_names=nomes, output_dict=True)
cla = pd.DataFrame(classific).transpose()
from sklearn.metrics import confusion_matrix
tn, fp, fn, tp = confusion_matrix(teste1, preds).ravel()
specificity = tn / (tn+fp)
Sensitivity = tp / (fn + tp)
```

Figura 16: Métricas de Avaliação de Desempenho. Fonte: o autor (2020).

4 AVALIAÇÃO

Para avaliar a classificação foi utilizado diferentes artifícios, um deles foi a matriz de confusão, esta é uma forma de expressar quantas das classificações estão corretas, quantas estão incorretas e em qual o classificador fica confuso. Em uma matriz de confusão as linhas representam as classes corretas enquanto as colunas representam a classificação gerada pelo algoritmo, a diagonal representa os números de acertos que deste algoritmo, ou seja, quando o classificador atribui corretamente a classe, os valores que saíram dessa diagonal representam

os erros.

	Predição	
Atual	Verdadeiro Positivo	Falso Negativo
	Falso Positivo	Verdadeiro Negativo

Tabela 8: Saída de dados Matrix de Confusão. Fonte: Autoria do autor.

Em seguida um exemplo da saída de dados

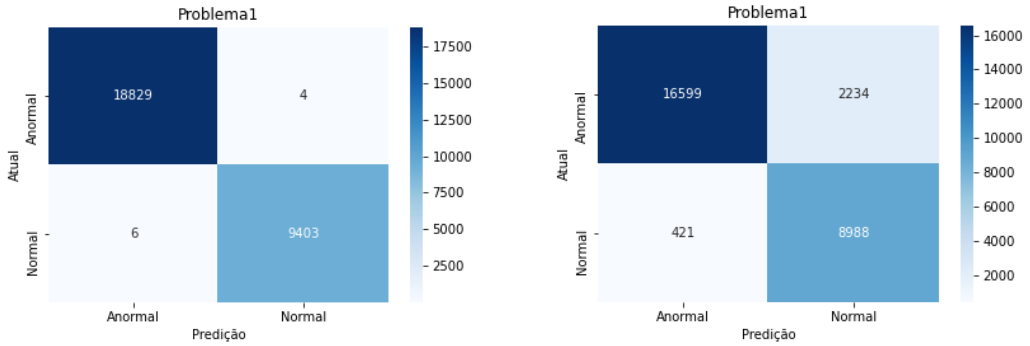

Figura 7: Matriz Confusão. Fonte: Autoria do autor.

Para obter uma melhor visão do desempenho deste modelo é utilizado métricas como *precision, recall, Specificity* e F1 *Score.*

Precision é a divisão do número de acertos de uma classe dividido pelo número total de vezes que aquele algoritmo classificou como sendo a mesma. (MISHRA, 2018)

$$Precision = \frac{Positivo\ Verdadeiro}{Verdadeiro\ Positivo + Falso\ Positivo}$$

O *sensitivity (recall)* é a medida da proporção de casos positivos que foram corretamente classificados como positivo, o cálculo dessa métrica é Positivo verdadeiro dividido pela quantidade total de itens que são positivos. Em resumo essa métrica refere-se à probabilidade do verdadeiro ser classificado como verdadeiro.

$$Recall = \frac{Positivo\ Verdadeiro}{Positivo\ Verdadeiro + Falso\ Negativo}$$

Specificity é a medida da proporção de casos negativos que foram corretamente classificados como negativo, o cálculo dessa métrica é negativo verdadeiro dividido pela quantidade total de itens que são negativos. Em resumo essa métrica refere-se à probabilidade do não verdadeiro ser classificado como não verdadeiro.

$$Specificity = \frac{\text{Negativo Verdadeiro}}{Negativo\ Verdadeiro + Falso\ Positivo}$$

Sensitivity e *Specitivity* tem origens na avaliação de testes de doença, ambas são medidas estratégicas de performance de classificação binaria.

O F1 *Score* é menos intuitivo que os anteriores pois este combina os dois anteriores em uma única métrica, se ambos os anteriores apresentarem um valor alto ou ambos baixo, este acompanha, caso um tenha valor alto e outro baixo este apresenta valor baixo, desta forma ele apresenta se o classificador realmente é bom em identificar as classes (KREIGER, 2020).

$$F1 = 2 * \frac{1}{\frac{1}{Precision} + \frac{1}{recall}}$$

RF		SVM	
F1:	99,96%	F1:	90,77%
Specificity:	99,98%	Specificity:	88,14%
Sensitivity:	99,94%	Sensitivity:	95,53%
Precision:	99,96%	Precision:	91,72%

Tabela 9: Resultados das métricas de avaliação. Fonte: o autor (2020)

5 ANÁLISE E DISCUSSÃO DE RESULTADOS

Para realizar a análise dos algoritmos, o presente estudo foi dividido em 2 experimentos:

(1) - A primeira etapa consiste em aglutinar os dados de todos os poços de petróleo que apresentam os mesmos problemas e aplicar os dois algoritmos, RF e SVM.

(2) O segundo experimento consiste em analisar individualmente cada poço/problema, para desta forma analisar o resultado do algoritmo na classificação dos mesmos.

5.1 EXPERIMENTO 1

Nesse sentido no primeiro experimento foram abordados os problemas descritos no capitulo 3, onde todos os poços que apresentavam o mesmo tipo de problema foram aglutinados em um único arquivo.

5.2 BSW

Para o problema de BSW o algoritmo RF apresentou um maior número de acertos comparado ao SVM, como podemos verificar na figura 17.

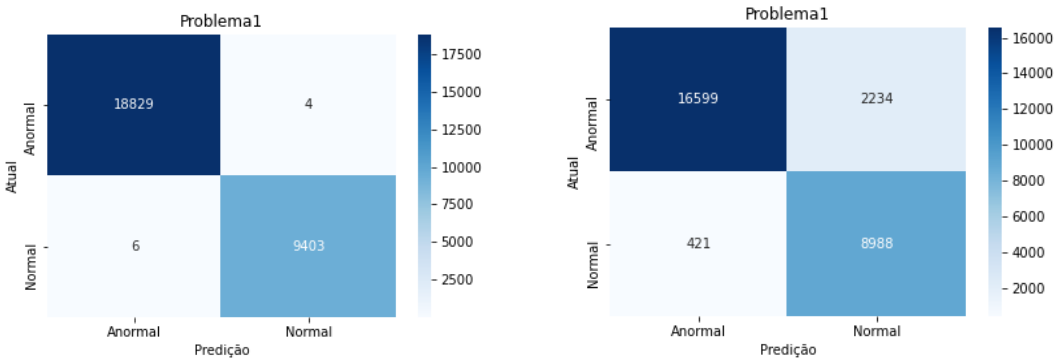

Figura 17: Matriz Confusão problema BSW. Fonte: o autor (2020).

Baseando nas métricas de avaliação utilizadas, o RF foi superior que o SVM para o problema em questão, apresentando dominância em todas as métricas avaliadas, como podemos verificar na tabela 10.

	RF		SVM
F1:	99,96%	F1:	90,77%
Specificity:	99,98%	Specificity:	88,14%
Sensitivity:	99,94%	Sensitivity:	95,53%
Precision:	99,96%	Precision:	91,72%

Tabela 10: Resultado das métricas de avaliação para o problema BSW. Fonte: o autor (2020)

5.2.1 DHSV

Seguindo a linha do problema anterior, na figura 18, podemos verificar que o DHSV apresentou resultado semelhante, no qual o RF foi mais assertivo, apresentando métricas ligeiramente superiores ao RF, estas identificadas na tabela 11.

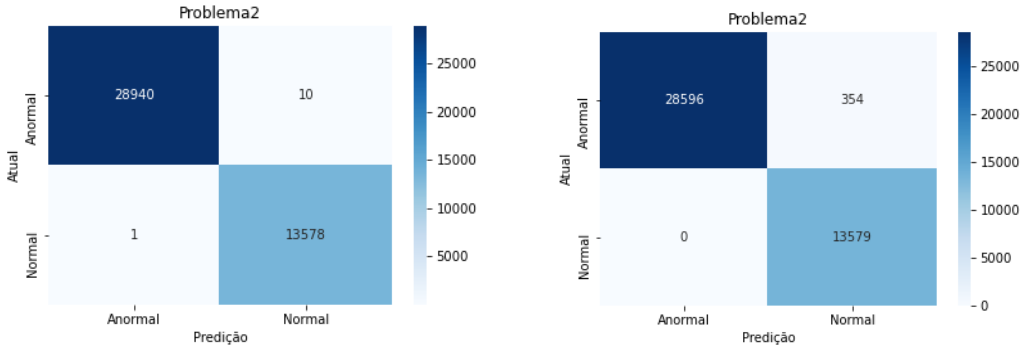

Figura 18: Matriz Confusão problema Fechamento DHSV. Fonte: o autor (2020).

RF		SVM	
F1:	100%	F1:	99,17%
Specificity:	100%	Specificity:	98,78%
Sensitivity:	100%	Sensitivity:	100,00%
Precision:	100%	Precision:	99,19%

Tabela 11: Resultado das métricas de avaliação para o problema fechamento DHSV. Fonte: o autor (2020)

5.2.2 Perda Repentina de Produtividade

Para o problema de perda repentina de produtividade, apesar das métrica F1, *Specificity* e *precision*, apresentarem melhores resultados para o RF, o SVM apresentou uma melhor *sensitivity* para o SVM, como pode ser observado na tabela 12.

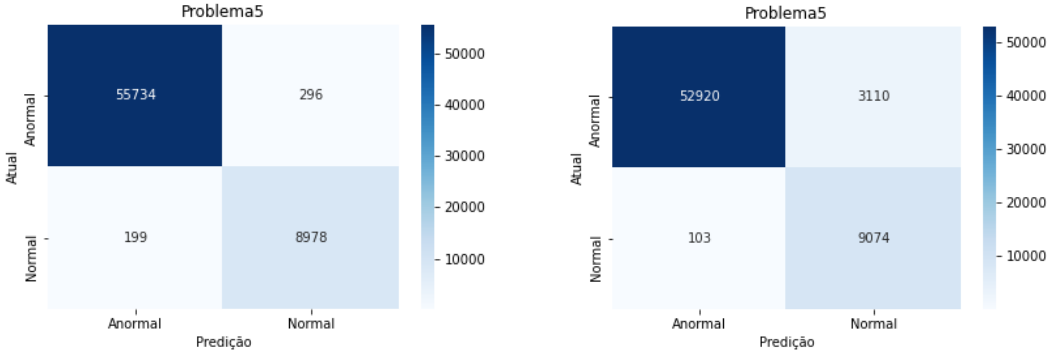

Figura 19: Matriz Confusão problema perda repentina de produtividade. Fonte: o autor (2020)

	RF		SVM
F1:	99,24%	F1:	95,35%
Specificity:	99,47%	Specificity:	94,45%
Sensitivity:	97,83%	Sensitivity:	98,88%
Precision:	99,25%	Precision:	96,24%

Tabela 12: Resultado das métricas de avaliação para o problema perda repentina de produtividade. Fonte: o autor (2020)

5.2.3 Hidrato

No caso do hidrato, o RF apresentou métricas melhores ou iguais o SVM, como podemos identificar na tabela 12.

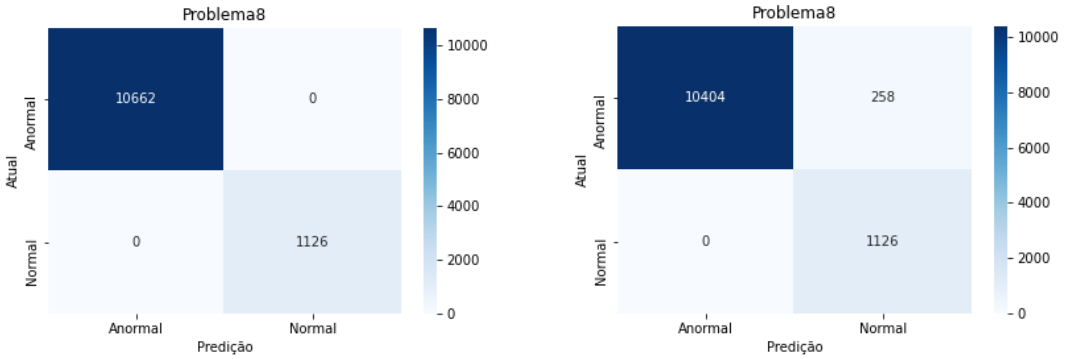

Figura 20: Matriz Confusão problema hidrato. Fonte: o autor (2020)

	RF		SVM
F1:	100%	F1:	97,91%
Specificity:	100%	Specificity:	97,58%
Sensitivity:	100%	Sensitivity:	100,00%
Precision:	100%	Precision:	98,22%

Tabela 13: Resultado das métricas de avaliação para o problema hidrato Fonte: o autor (2020)

5.2.4 Experimento 2

Para o segundo bloco de experimentos foi aglutinado em um arquivo todos os poços analisados no item anterior, para desta forma avaliar a capacidade do algoritmo de identificar anomalias dentro do conjunto de dados a partir do treinamento do conjunto total de dados, não estando separados por problema.

Mesmo com o treinamento de todos os problemas em um único conjunto de treino, o algoritmo se mostrou capaz de classificar corretamente as classes, porém o RF se demostrou, mais uma vez, ser superior ao SVM para o tipo de classificação abordado.

Conjunto de problemas

RF SVM

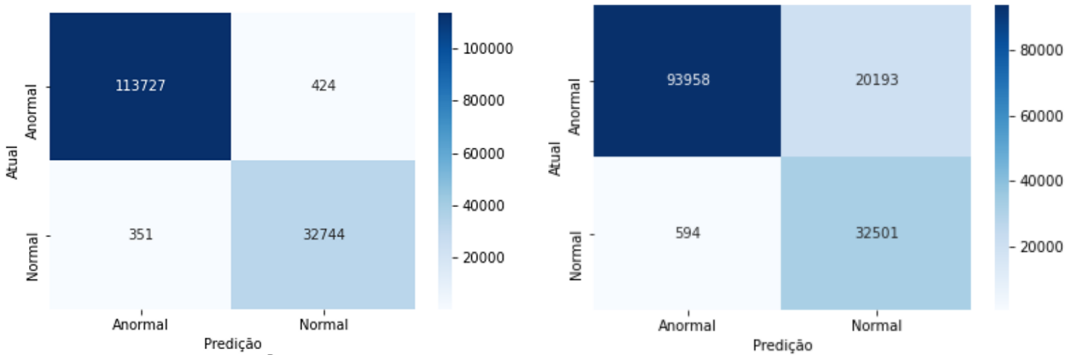

Figura 21: Matriz confusão para conjunto de problemas. Fonte: o autor (2020).

	RF		SVM
F1:	99,47%	F1:	86,83%
Specificity:	99,63%	specificity:	82,31%
Sensitivity	98,94%	Sensitivity	98,21%
Precision:	99,47%	Precision:	90,90%

Tabela 14: Resultado das métricas de avaliação para conjunto de problemas. Fonte: o autor (2020)

No primeiro bloco de experimentos o RF apresentou resultados superiores para todas as métricas de avaliação do algoritmo exceto *sensitivity* para o problema de perda repentina de produtividade, desta forma ele foi identificado como o melhor que o SVM para classificar o problema.

No segundo bloco de experimentos ao aglutinar todos os problemas o algoritmo teve um desempenho inferior comparado a média do desempenho ao realizar o treino e a classificação com os arquivos separados por problema. Desta forma, a utilização de um conjunto de treino onde os dados são separados de acordo com o problema classificado aumentou a acurácia do modelo.

O conjunto de dados utilizados se apresentaram muito comportados nos indicando que seriam dados sintéticos, pois dados reais dificilmente apresentam variáveis tão comportadas, o alto desempenho dos algoritmos indicados pelas métricas de avaliação podem ser alterados negativamente com um conjunto de dados real, sendo necessário uma nova avaliação.

Os dados utilizados nessa analise foram limitados a algumas das inúmeras variáveis que se tem em um conjunto de exploração, desta forma esse estudo poderia ser feito de forma mais abrangente para englobar um maior número de sensores, de fundo tanto quanto de superfícies além das válvulas presentes no conjunto de exploração.

O estudo foi delimitado a classificar 4 problemas dentre os vários presentes na exploração de exploração de petróleo de águas profundas, os problemas trados foram classificados como normal e anormal, considerando a classe de falha presente no conjunto de dados como um estado anormal, outro estudo poderia ser executado com a finalidade de classificar as classes em cada um dos problemas, não se limitando a identificar o estado de produção normal.

6 CONCLUSÕES

A pesquisa desenvolvida neste trabalho de conclusão de curso aborda a utilização de ferramenta de aprendizado de máquina a fim de diferenciar problemas de produção como Perda Repentina de Produtividade, Hidrato, BSW, e DHSV a partir de dados reais disponíveis na literatura. Com base no comportamento de variáveis localizadas na árvore de natal foi possível analisar se o poço apresentava condições normais de produção.

Para detectar a presença de anomalias na linha de produção foi utilizado duas técnicas de aprendizado de máquina, RF e SVM, ambos apresentaram bons resultados de classificação, de acordo com as métricas *precision, recall, specificity e F1 Score.*

6.1 RESPOSTAS ÀS QUESTÕES DE PESQUISA

A partir da aplicação das técnicas descritas no capitulo 2, os resultados obtidos foram condizentes com as expectativas, de acordo com as métricas *precision, recall, specificity e F1 Score,* utilizadas para avaliar o desempenho das técnicas utilizadas. Salienta-se que foi identificado um melhor desempenho do RF na classificação destes eventos indesejados, a ferramenta se mostrou eficaz. Desta forma, foi verificado que as variáveis presentes na árvore de natal que foram utilizadas nesse estudo são suficientes para a classificação de estado da linha de produção, classificando-as, corretamente, quando o sistema apresenta condições normais de produção.

Para se classificar fenômenos indesejados pode-se desenvolver um algoritmo para classificar em tempo real a entrada de dados diretas do sistema para identificar qual o estado atual dos da linha de produção. Ao treinar o algoritmo com os dados de determinado problema, os dados recebidos em tempo real podem ser classificados pela ferramenta.

6.2 DESDOBRAMENTOS FUTUROS DA MODELAGEM DESENVOLVIDA

Uma das questões a serem colocadas como continuidade e desdobramento do estudo seria a aplicação das ferramentas estudas em um conjunto de dados reais, uma vez que no ambiente de produção muitas vezes não se tem a qualidade de dados como o apresentado no conjunto de dados utilizado. Desta forma, seria analisada a capacidade preditiva e a utilização da ferramenta em ambiente de produção, uma vez que os dados recebidos ocorrem em tempo

real e o algoritmo teria um determinado intervalo de tempo para informar o operador sobre a alteração de estado e, assim, o mesmo teria um tempo hábil para realizar manobras para diminuir ou evitar efeitos negativos do problema na linha de produção.

Como analisado no capitulo 3, o conjunto de dados apresenta 8 variáveis coletadas ao longo da linha de produção. Por meio desses dados, a utilização de combinações de variáveis para a aplicação da ferramenta poderia ser um novo desdobramento do estudo, tendo em vista a qualidade das variáveis no momento analisado. Além do desdobramento citado, a utilização de variáveis distintas poderia ensejar outros estudos comparativos de técnicas de aprendizado de máquina, como Redes Neurais, *Adaboost e Naive Bayes*. Portanto, a IA é um campo de estudos promissor que deve ser explorado a fim de solucionar os mais variados problemas de produção.

REFERÊNCIAS

ABASS, H.; BASS, D. **The critical production rate in water - Coning system**. PermianBasin Oil and Gas Recovery Conference. Texas: [s.n.]. 1988. p. 351-360.

ANDREOLLI, I. **Introdução à Elevação e Escoamento Monofásico e Multifásico de Petróleo**. Rio de Janeiro: Interciência, 2016.

ANP. Boletim da Produção de Petróleo e Gás Natural – Circulação Externa, Brasília, p. 32, 2019. Disponivel em: <http://www.anp.gov.br/images/publicacoes/boletins-anp/Boletim_Mensal-Producao_Petroleo_Gas_Natural/boletim-janeiro-2019.pdf>. Acesso em: 23 Agosto 2020.

BAKER, R. **A Primer of Offshore Operations**. 3ª. ed. Texas: Petroleum Extension Service, 1998.

BREIMAN, L. Random forests. **Machine learning**, 45, 2001. 5–32.

BREIMAN, L. et al. **Classification and Regression Trees**. Belmont: Routledge, 2017. 368 p.

CENTRO BRASILEIRO DE INFRAESTRUTURA. Quantos poços de petróleo e gástemos no Brasil? **CBIE**, 2019. Disponivel em: <https://cbie.com.br/artigos/quantos-pocos-de-petroleo-e-gas-temos-no-brasil/>. Acesso em: 23 Agosto 2020.

CHAKURE, A. Random Forest Regression. **Towards datascience**, 23 Agosto 2019. Disponivel em: <https://towardsdatascience.com/random-forest-and-its-implementation-71824ced454f>.

CHAPMAN, P. et al. CRISP-DM 1.0: Step-by-step data mining guide. **CRISP**, 2000. Disponível em: <https://www.the-modeling-agency.com/crisp-dm.pdf>. Acesso em: 23 Agosto 2020.

DYLAN. The Data Mining Process (CRISP-DM). **Nimble Coding**, 2016. Disponível em: <https://www.nimblecoding.com/data-mining-process-crisp-dm/>. Acesso em: 23 Agosto 2020.

FACELI, K. et al. **Inteligência Artificial:** Uma Abordagem de Aprendizado de Máquina. 1ª. ed. São Paulo: LTC, 2011. 394 p.

FREITAS, A. Árvores de Decisão. **ulisboa**, 2004. Disponível em: <http://web.tecnico.ulisboa.pt/ana.freitas/bioinformatics.ath.cx/bioinformatics.ath.cx/indexf23d.html?id>. Acesso em: 23 Agosto 2020.

GOMES, J. S.; BARATA, F. **O Universo da Indústria Petrolífera.** 3ª. ed. Lisboa: Calouste Gulbenkian, 2007. 647 p.

HAUSLER, R. H.; KRISHNAMURTHY, R. M.; SHERAR, B. W. A. Observation of Productivity Loss in Large Oil Wells due to Scale Formation without Apparent Production of

Formation Brine. **NACE International**, Dallas, Março 2015.

KOEHRSEN, W. Random Forest Simple Explanation. **Medium**, 2020. Disponível em: <https://medium.com/@williamkoehrsen/random-forest-simple-explanation- 377895a60d2d>. Acesso em: 23 agosto 2020.

KOTSIANTIS, S. B.; ZAHARAKIS, I. D.; PINTELAS, P. E. Machine learning: a review of classification and combining techniques. **Artificial Intelligence Review**, 26, 10 Novembro 2007. 159-190.

KREIGER, J. Evaluating a Random Forest model. **Analytics Vidhya**, 2020. Disponível em: <https://medium.com/analytics-vidhya/evaluating-a-random-forest-model-9d165595ad56>. Acesso em: 23 Agosto 2020.

MEGLIO, F. et al. Stabilizationof slugging in oil production facilities with or without upstream pressure sensors. J.Process Control 22 (4), 809–822. **Journal of Process Control**, 22, Abril 2012. 809-822.

MICHALSKI, R. S.; CARBONELL, J. G.; MITCHELL, T. M. **Machine learning:** An Artificial Intelligence Approach. [S.l.]: Springer, 2013. 572 p.

MISHRA, A. Metrics to Evaluate your Machine Learning Algorithm. **Towards data science**, 2018. Disponivel em: <https://towardsdatascience.com/metrics-to-evaluate-your-machine-learning-algorithm-f10ba6e38234>. Acesso em: 05 dez. 2020.

MITCHELL, T. M. **Machine Learning**. Nova York: McGraw Hill, 1997. 414 p.
NASRABADI, N. M. Pattern recognition and machine learning. Journal of electronic. **Journal of Electronic Imaging**, 16, 1 Outubro 2007.

PEDREGOSA, F. et al. Scikit-learn: Machine Learning in Python. **Journal of Machine Learning Research**, Massachusetts, v. 12, p. 2825-2830, 2011. ISSN ISSN 1533-7928. Disponivel em: <https://scikit-learn.org/stable/about.html#citing-scikit-learn>. Acesso em: 06 dez. 2020.

PIERRE, D. **Essentials of Reservoir Engineering**. Paris: Editions Technip, 2007.

PODOROZHNIKOV, S. et al. Dielectric Method for Diagnosing Formation of Gas Hydrates in Gas Pipelines. **MATEC Web of Conferences**, 73, 2016. 7.

QUINLAN, J. R. Induction of decision trees. **Machine Learning**, março 1986. 81–106.

REZENDE, S. O. **Sistemas inteligentes:** fundamentos e aplicações. Barueri: Manole, 2003. 525 p.

RUSSELL, S. J.; NORVIG, P. **Artificial intelligence:** a modern approach. 3ª. ed. New Jersey: Prentice Hall, 2010. 1152 p.

SANTOS, Mayara de Jesus Rocha. **DETECÇÃO DE PROBLEMAS DE GARANTIA DE ESCOAMENTO A PARTIR DA UTILIZAÇÃO DE FERRAMENTAS DE APRENDIZADO DE MÁQUINA**. 2020. 110 f. Dissertação (Mestrado) - Curso de

Programa Francisco Eduardo Mourão Saboya de Pós-Graduação em Engenharia Mecânica, Engenharia de Mecânica, Universidade Federal Fluminense, Niterói, 2020.

SANTOS, V. C. P. D. et al. Hidratos em perfurações de poços com elevação de bombeio por cavidades progressivas. **Cadernos de Graduação: Ciências exatas e tecnológicas**, Alagoas, v. 6, n. 1, p. 19-34, abril 2020.

SCHLUMBERGER. **The Schlumberger Oilfield Glossary**, 2020. Disponível em: <https://www.glossary.oilfield.slb.com/>. Acesso em: 23 Agosto 2020.
SCHMIDT, Z.; DOTY, D.; DUTTA-ROY, K. Severe slugging in offshorepipeline riser-pipe systems. **Society of Petroleum Engineers Journal**, 25, Fevereiro 1985.

SOUSA, R. G. História do Petróleo no Brasil. **Brasil Escola**, 2020. Disponível em: <https://brasilescola.uol.com.br/brasil/historia-do-petroleo-no-brasil.htm.>. Acesso em: 22 Agosto 2020.

TAKEI, J. et al. Flow Instability In Deepwater Flowlines And Risers - A Case Study Of Subsea Oil Production From Chinguetti Field, Mauritania. **Society of Petroleum Engineers**, Janeiro 2010.
THEYAB, M. A. Severe Slugging Control: Simulation of Real Case Study. **Journal of Environmental Research**, Londres, 2, 27 Março 2018.

THOMAS, J. E. **Fundamentos de Engenharia de Petróleo**. Rio de Janeiro: Interdência, 2001.

TURING, A. M. Computing machinery and intelligence. **Mind**, Manchester, LIX, 1 Outubro 1950. 433-460.

U.S. ENERGY INFORMATION ADMINISTRATION. **International Energy Outlook 2019 with projections to 2050**. September 24, 2019. Washington, p. 85. 2019.

VARGAS, R. E. V. et al. The first realistic and public dataset with rare undesirable real events in oil wells. **Journal of Petroleum Science and Engineering**, Julho 2019.

WIRTH, R.; HIPP, J. CRISP-DM: Towards a Standard Process Model for Data. **Proceedings of the 4th International Conference on the Practical Applications of Knowledge Discovery and Data Mining.** , , Janeiro 2000. 11.

YADAV, A. SUPPORT VECTOR MACHINES(SVM). **Towards data science**, 2020. Disponivel em: <https://towardsdatascience.com/support-vector-machines-svm-c9ef22815589>. Acesso em: 06 dez. 2020.