Universidade Presbiteriana Mackenzie

Pró-Reitoria de Extensão e Educação Continuada

Faculdade de Computação e Informática

Pós-Graduação *Lato Sensu*

Identificação de anomalia no processo de extração de petróleo em poços marítimos por meio de aprendizado de máquina

Lucas Defante

São Paulo

2024

Universidade Presbiteriana Mackenzie

Lucas Defante

Identificação de anomalia no processo de extração de petróleo em poços marítimos por meio de aprendizado de máquina

Trabalho de Aplicação de Conhecimento apresentado ao Curso de Pós-Graduação em Inteligência Artificial da Universidade Presbiteriana Mackenzie para a obtenção do título de Especialista em Inteligência Artificial.

Orientadora: Profa. Dra. Pollyana Coelho da Silva Notargiacomo

São Paulo

2024

Resumo

Durante o ano de 2023, 80% da produção de petróleo da Petrobras se deu em plataformas *offshore* de extração e problemas associados a esse tipo de atividade industrial podem tomar uma boa parte dos investimentos destinados, pois demandam alto custo de manutenção e mão de obra especializada, além dos possíveis impactos gerados na própria imagem da empresa em acidentes causados devido a esses problemas. Com isso em vista, este trabalho propôs-se a criar classificadores para identificação de apenas um dos tipos dos problemas associados, a anomalia instabilidade do fluxo. Através do projeto 3W, uma iniciativa da Petrobras, foi possível extrair dados de observações de diversos poços de produção da empresa, onde diversas anomalias são identificadas, bem como as de funcionamento normal. Foram treinados 3 modelos de classificadores: *Random Forest*, *Naive Bayes* e *Nearest Neighbors*, onde todos eles obtiveram uma acurácia acima de 90%, indicando que podem ser utilizados em ambientes reais de observação dos poços para identificação da anomalia estudada com o ganho da possibilidade de ação rápida para correção e evitar a evolução da situação para problemas mais críticos.

Palavras-chave: *Random Forest*. *Naive Bayes*. *Nearest Neighbors*. Algoritmo de classificação. Poços de petróleo. Identificação de anomalia.

SUMÁRIO

LISTA DE QUADROS

LISTA DE FIGURAS

LISTA DE GRÁFICOS

INTRODUÇÃO

A produção de petróleo envolve muitos riscos devido às extremas condições em que são submetidos os equipamentos, algumas dessas condições envolvem alta pressão, temperaturas baixas e tudo isso longe do continente em plataformas no oceano adentro o que dificultam qualquer tipo de atuação para manutenção ou mesmo calibração de sensores submergidos próximos à válvula de extração.

No estudo Vargas et al. (2019) foi divulgado algumas das principais anomalias associadas a esse processo de extração:

1. Aumento Abrupto de BSW (*Basic Sediment and Water*)
2. Fechamento Espúrio de DHSV (*Downhole Safety Valve*)
3. Intermitência Severa
4. Instabilidade no Fluxo
5. Perda Rápida de Produtividade
6. Restrição Rápida em CKP (*Choke* de Produção)
7. Incrustação em CKP
8. Hidrato em Linha de Produção

Todos esses eventos são indesejados e podem representar algum atraso na produção do óleo ou mesmo atuações mais invasivas no sistema para o restabelecimento das condições normais de trabalho.

É nesse momento que a inteligência artificial ganha potência. Ao saber identificar as anomalias indesejadas no processo, ganha-se a capacidade de prever quando alguma situação não está conforme esperada e pode desencadear ações para rápida remediação ou completa evitação de um problema em potencial.

No presente estudo o foco está na identificação e predição de eventos anômalos do tipo número 4 – Instabilidade no Fluxo. A escolha por essa classe se deu pelo equilíbrio entre eventos reais identificados da anomalia e eventos considerados normais. Assim é possível deixar de lado outras linhas de estudo que são completamente dedicadas à problemas que tratam do desbalanceamento de classes e eventos artificiais gerados para suprir e equilibrar essa diferença.

A Instabilidade no Fluxo pode evoluir para uma anomalia do tipo Intermitência Severa,

um tipo muito mais grave para ser remediado (VARGAS et al., 2019). Portanto, sua rápida identificação pode ser determinante na saúde do processo como um todo.

Por fim, serão comparados performances e desempenhos de diferentes modelos de aprendizado de máquina que tratam da classificação binária, no nosso contexto: i – classe normal; ii – classe anômala instabilidade no fluxo.

1 RELATÓRIO DA SITUAÇÃO

Aqui serão apresentados alguns pontos que dizem respeito ao contexto do problema, a empresa, os objetivos do estudo e os sintomas associados.

1.1 A empresa

A Petrobrás, uma empresa de capital aberto sendo o governo federal o acionista majoritário, é referência no âmbito da exploração e extração de petróleo, tendo monopolizado o mercado brasileiro até 1997 desde sua criação em 1953.

Apenas no ano de 2023, foram produzidos pela empresa 2.748 mil barris de petróleo bruto e gás natural por dia, sendo que quase 80% da produção foi realizada em estações *offshore*.

Também foi responsável pela descoberta e exploração da camada pré-sal, que segundo a empresa, representa 1/3 da produção em toda a América Latina e com toda sua receita impulsiona desenvolvimento e fomentação da economia brasileira.

1.2 Desafio

A abordagem desse problema será a utilização de um modelo de classificação de aprendizado de máquina para atribuir eventos aos tipos normal e anomalia instabilidade no fluxo. Os dados utilizados são da contribuição feita pelo estudo Vargas et al. (2019) denominado projeto 3W da Petrobrás onde diversos poços foram monitorados e classificados entre os diversos tipos de anomalias encontrados no processo de extração de petróleo.

O projeto 3W tem como objetivo a estimulação, experimentação e desenvolvimento de projetos de aprendizado de máquina ao abordar problemas de classificação específicos em poços de extração de petróleo.

1.3 Os sintomas

Até recentemente todo processo de identificação de anomalias no processo de extração é feito a partir da cognição humana, onde um especialista observa os dados de monitoração e consegue apontar se existe algum problema e de qual natureza (VARGAS et al., 2019).

Como qualquer processo manual, ele demanda tempo e mão de obra especializada, muitas vezes o diagnóstico pode ser tardio limitando as formas mais preventivas e corretivas de

tratar ou evitar alguma anomalia.

A implementação de um processo inteligente e automatizado permite a rápida identificação e tratamento com um grau de eficácia satisfatório para as necessidades da situação.

1.4 Objetivo

Este trabalho tem como objetivo avaliar a performance de alguns algoritmos de inteligência artificial aplicado a um problema de classificação binária.

Ao final, será proposta uma solução de modelo que consiga com acurácia satisfatória a identificação da anomalia estudada através dos dados dos poços observados. Uma solução desse tipo permite a monitoração e classificação em tempo real de problemas em plataformas de extração, possibilitando uma rápida identificação e atuação pelas equipes responsáveis, evitando assim possíveis desastres naturais e não impactando a produção.

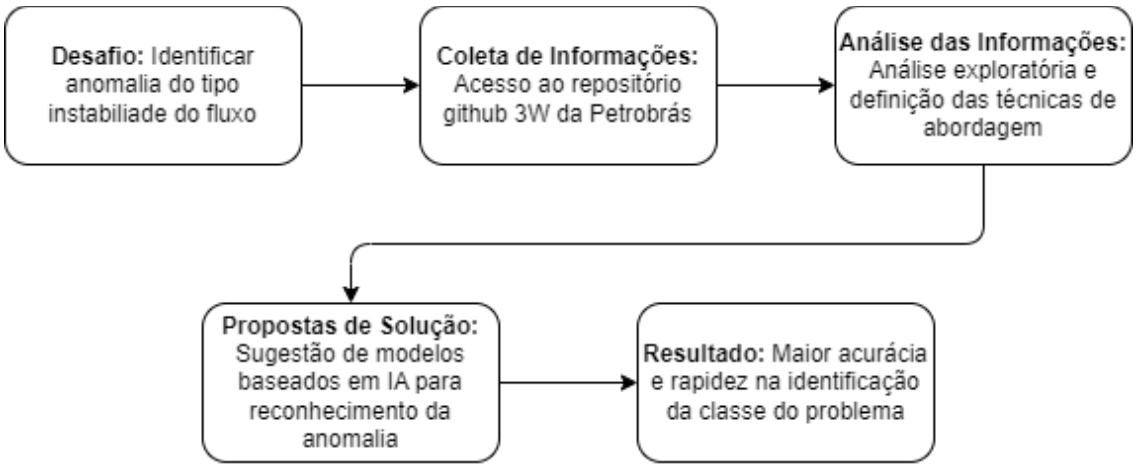
Também a contribuição a nível acadêmico da abordagem através de inteligência artificial de um problema real e extremamente relevante no Brasil e no mundo. Promoção da discussão e experimentação de diferentes formas de tratamento de dados no treinamento dos modelos.

2 DIAGNÓSTICO

Nesse capítulo será explicado a forma com que os dados serão tratados, a abordagem da análise e proposta de solução será executada.

A figura abaixo representa a gestão do tratamento do problema em questão e todas as fases que compõem seu entendimento, proposta de solução e avaliação de seu impacto.

Figura 1 – Diagnóstico Organizacional



Fonte: Elaborada pelo autor.

Para definição do desafio foi necessário o estudo de pesquisas efetuadas no campo relacional do tema, assim delimitando qual objetivo será buscado. Em seguida, para a coleta de informações deve-se realizar o acesso aos dados através do repositório do *dataset* disponível em https://github.com/petrobras/3W que foi publicado pelo perfil oficial da empresa na ferramenta.

Em seguida é feito uma análise exploratória para avaliar a representatividade da anomalia instabilidade do fluxo quando comparada com a situação normal de operação da extração e assim ponderar e contextualizar os resultados obtidos posteriormente no treinamento dos modelos.

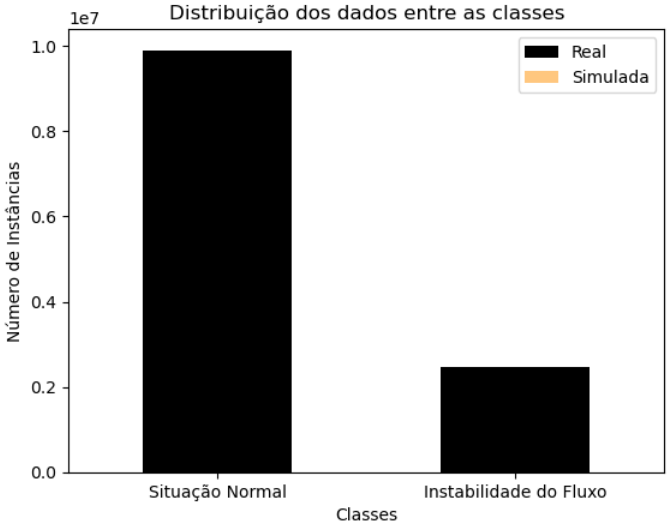
Após a identificação da distribuição dos dados disponíveis é o momento onde pode-se aplicar uma análise quantitativa com o objetivo de definir quais características dos dados são relevantes para a utilização no modelo de treinamento. Como apontado por Vargas et al. (2019) muitos sensores por vezes ficam congelados ou falham e não apresentam valores dado as condições extremas em que se encontram em operação, sendo assim acabam sendo irrelevantes na distinção das classes entre as instâncias, possibilitando a geração de um modelo mais leve e mais rápido.

2.1 As informações

Pelo repositório 3W (Anexo A) são disponibilizadas 9,9 milhões de instâncias classificadas como situação normal de operação sendo 100% de origem real das coletas dos sensores e 2,4 milhões de instâncias classificadas como anomalia do tipo instabilidade do fluxo sendo 100% de origem real. Não existem dados simulados para esse tipo de anomalia, portanto podemos esperar alguns efeitos do desbalanceamento entre as classes na performance dos modelos que serão construídos, esse problema bastante conhecido na área da inteligência artificial pode ser um fator negativo na performance quando apresentado de forma acentuada.

Existem diversos estudos acerta de técnicas aplicadas para tratamento desse tipo de situação, alguns desses métodos envolvem utilização de métricas de avaliação apropriadas que levam em consideração casos raros, segmentação dos dados, aprendizado de somente dos casos raros, entre outros, aprofundados em Weiss (2004). Portanto foi estudado a aplicação de algumas dessas técnicas pertinentes para reduzir esse desequilíbrio e guiar o processo da modelagem das soluções propostas.

Gráfico 1 – Distribuição dos dados analisados



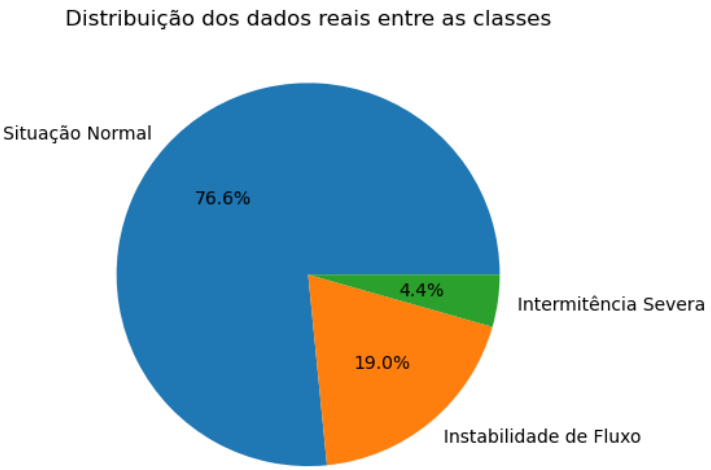
Fonte: Elaborada pelo autor

Conforme descrição da base de dados (Apêndice A), ela apresenta 10 colunas sendo a primeira com informação do horário de captura da informação dos sensores, a última com o valor nominal da classe daquela instância (normal, anomalia ou estado transitório) e o restante os valores capturados dos sensores no momento.

Através da análise exploratória dos dados observa-se que para ambas as classes algumas características não são relevantes na análise, portanto todas elas foram descartadas e não alimentadas no treinamento dos modelos. Descartou-se a coluna T-JUS-CKGL por não apresentar valor, a coluna *timestamp* também foi descartada pois apesar de trazer o momento no tempo em que a observação foi realizada, os dados já estão ordenados cronologicamente pelo *index*.

Um ponto importante na definição da análise do tipo de anomalia instabilidade no fluxo é que como já mencionado ele é um estado anterior da anomalia intermitência severa, um tipo de problema mais crítico que demanda a atenção necessária desse estado, portanto ao anunciar seu predecessor com maior rapidez e acurácia é possível evitar até aproximadamente 4% de estados em que esse tipo de anomalia mais sério foi identificado e aproximadamente 25% considerando ambas situações de anomalia conforme distribuição dos dados reais apresentado na Figura 3 abaixo.

Gráfico 2 – Distribuição dos dados reais



Fonte: Elaborada pelo autor

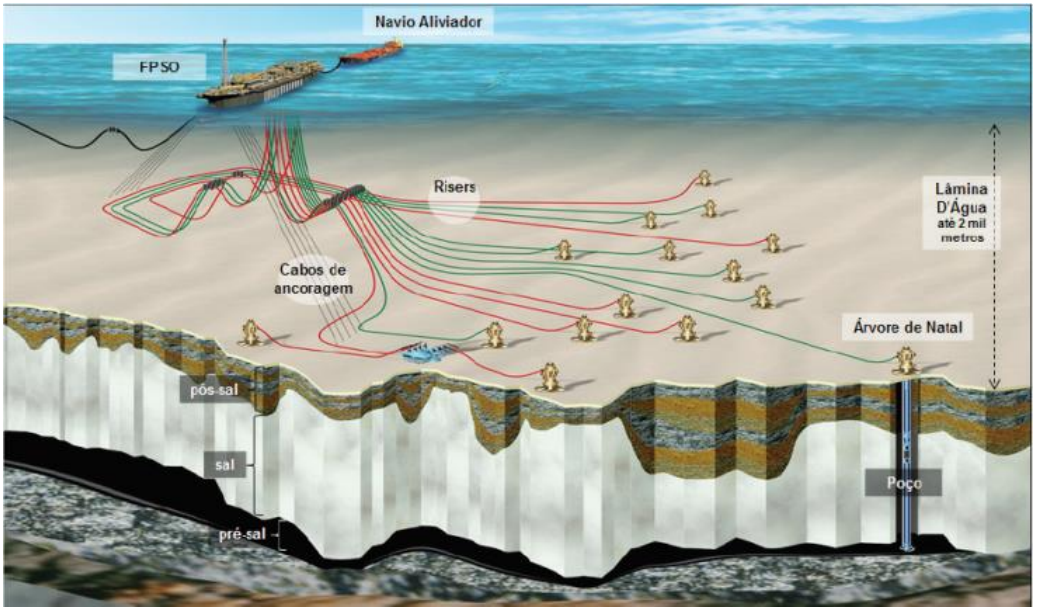
2.2 Análise e diagnóstico

Segundo Andreolli (2018), e ilustrado na Figura 4, uma estação de extração de petróleo em alto mar é composta de diversos poços de extração todos conectados à mesma plataforma e navio correspondente de forma a maximizar a produção naquela localidade.

O risco associado à algum acidente nessas regiões está diretamente relacionado aos tamanhos das áreas e quantidade de poços existentes de modo que se faz extremamente necessário a previsibilidade dessas situações indesejadas. Considerando os quase 25% de

representatividade das anomalias aqui em foco essa importância torna-se ainda maior quando colocamos em perspectiva os trabalhadores que se encontram nessas plataformas e o custo operacional de uma manutenção marítima dessa magnitude.

Figura 2 – Ilustração da exploração de petróleo em alto mar



Fonte: Andreolli, 2018

A intermitência severa se dá ao produzir gás seguido de líquido de forma cíclica e tem como efeito prático a alteração da pressão em componentes do sistema de produção (ANDREOLLI, 2018), portanto espera-se que sensores de pressão tenham uma importante influência na capacidade do modelo de identificar tanto a instabilidade no fluxo quanto a sua forma de evolução mais crítica.

No capítulo seguinte serão exploradas soluções que expliquem os dados apresentados anteriormente, de forma com que se conclua qual algoritmo é mais eficiente na distinção da anomalia, isso pode ser um fator primordial na manutenção dos sistemas de monitoração de produção de petróleo, permitindo um investimento mais eficiente e eventualmente alterar a forma da monitoração passando a ser derivada por evento disparado pelo modelo treinado para identificar a falha.

Além da identificação da instabilidade do fluxo é interessante e necessário uma outra abordagem, que não será perseguido pelo presente estudo, para que consiga relacionar as características das observações com os diferentes tipos de anomalias apresentadas pelos dados.

3 SOLUÇÃO

Esse capítulo terá como foco a apresentação de soluções para a identificação de anomalia do tipo instabilidade no fluxo, onde anteriormente foram analisados os dados referentes a esse tipo de evento com ocorrências em plataformas marítimas de produção de petróleo.

Na seção anterior foi destacado um desbalanceamento nos dados entre as classes normal e anomalia que poderia afetar o desempenho dos classificadores em eventos reais de observação, dessa forma, a proposta apresentada também considerou esse efeito e formas de contorná-la caso ocorra quando em utilização.

Por fim, são destacados características positivas e negativas da solução onde percebe-se um claro potencial de melhora se comparado com a atual situação de atuação nas anomalias apenas após confirmação através de validadores humanos, e também as técnicas utilizadas para classificação são validadas através de referenciais teóricos que relacionam com pontos da solução proposta.

3.1 Propostas de solução

➢ **Proposta A: Combinação de classificadores do tipo *Random Forest, Naive Bayes* e *Forest Ensemble* para identificar anomalias do tipo instabilidade no fluxo**

- **Descrição da proposta:** Consiste na combinação de 3 tipos de classificadores diferentes, mencionados anteriormente, com intuito da identificação em tempo real de uma anomalia do tipo instabilidade no fluxo. Valendo-se de uma infraestrutura tecnológica preparada para aquisição e monitoração dos dados em tempo real da extração do petróleo e profissional especializado na confirmação das previsões realizadas pelos modelos a fim de retroalimentar e permitir uma melhoria contínua dos modelos.

Podem-se valer também de variações da proposta, realizando comparações entre o desempenho dos classificadores de forma independente ou conjunta atuando com uma combinação da previsão:

i. Combinação das probabilidades dos 3 classificadores para definir se a observação é uma anomalia ou não
ii. Utilização independente da classificação do modelo tipo *Random Forest*
iii. Utilização independente da classificação do modelo tipo *Naive Bayes*
iv. Utilização independente da classificação do modelo tipo *Forest Ensemble*

- **Possíveis impactos:** Uma mudança na metodologia de atuação desses tipos de eventos claramente ocorrerá, pois para utilizar-se dos benefícios de uma previsão e classificação em tempo real, processos devem ser discutidos e implementados, como por exemplo a disposição de passos a serem executados após uma geração de alerta pelo modelo, pois antes de qualquer ação de interrupção ou atuação direta nos componentes de extração é necessário dispor de um profissional com conhecimento técnico para confirmar tal evento e orientar as ações de acordo a fim de evitar que a situação se agrave caso seja confirmada ou não ser super sensível à monitoração e reagir precipitadamente caso ocorra um falso positivo. Contrastando claramente com uma situação sem monitoração baseada em aprendizado de máquina onde após confirmação do evento deve-se lidar com seus desdobramentos possivelmente mais críticos.

- **Prós:** A principal mudança que a utilização de classificadores permite é a identificação da anomalia em tempo real, possibilitando uma atuação focada e especializada para evitar um aumento de criticidade do evento e consequentemente uma melhor gestão de riscos na produção de petróleo, o que traz consigo todos seus aspectos positivos tanto na questão financeira quanto na imagem da empresa que se associa cada vez menos com problemas.

- **Contras:** Uma possibilidade cujo aprofundamento não foi realizado nesse presente estudo é a classificação e geração de falsos positivos anômalos nas observações, dado que para treinamento dos modelos foram utilizados os dados referentes apenas às classes normal e anomalia do tipo instabilidade no fluxo, desconsiderando observações geradas artificialmente e de outros tipos de anomalias já mencionados. Isso pode gerar um excesso de alertas, interrupções indevidas da produção e acionamento do profissional dedicado a verificar esse tipo de evento.

- **Recursos:** O recurso mais demandado nessa proposta é a pessoa especialista em inteligência artificial que deverá atuar em conjunto com uma pessoa engenheira de software, onde a primeira terá a incumbência de realizar a modelagem e entrega dos classificadores e a segunda ficará encarregada com a integração dos classificadores com o sistema de monitoração em tempo real dos sensores da plataforma de produção em infraestrutura tecnológica já em uso para esse objetivo. Dispondo desses recursos, o profissional responsável da plataforma ou setor designado terá a capacidade para agir de acordo após disparo dos alertas configurados.

Atualmente, o especialista em inteligência artificial recebe em média seis mil e cem reais por mês e o engenheiro de software seis mil e trezentos reais mensais, sendo esses os

únicos valores a serem investidos pela empresa.

- **Teoria de suporte e autor:**

No estudo Li (2022) é apresentada uma abordagem para lidar com a identificação de falhas em máquinas rotacionais de larga escala utilizando uma combinação de Redes Neurais Profundas e *Random Forests*. Nele, concluiu-se que com a utilização deste último algoritmo na solução final do modelo apresentado a acurácia na identificação apresentou melhora quando comparada ao estado da arte de classificadores do tipo Redes Neurais Profundas. *Random Forest* também é reconhecido por ser robusto em problemas de classificação por possuir uma sólida base nos princípios matemáticos e estatísticos (BHATTACHARYA, 2018).

Em Amin (2020) o algoritmo *Naive Bayes* foi utilizado para detecção de falha e diagnóstico em processos da indústria química. Como resultado de sua utilização é a geração de probabilidades associadas a cada classe de falha existente no problema relatado, o que permite o alerta nos determinados processos e atuação de acordo dado sua classe. Essa abordagem tem uma clara vantagem positiva ao possibilitar o funcionamento seguro e sem eventos da fábrica.

A utilização do algoritmo *Nearest Neighbors* como base da solução para identificação de falhas e diagnóstico em sistemas de geração de energia a partir de células fotovoltaicas foi realizada diversas vezes. Uma proposta de solução desenvolvida em Mouleloued (2023) tendo como base esse algoritmo obteve performance notavelmente superior quando comparada com outros algoritmos considerando as métricas de avaliação acurácia, precisão, *recall* e tempo de execução. Também em Harrou (2019) aproveitou-se de características do *Nearest Neighbors* para supervisionar esses sistemas de geração fotovoltaicas e obteve-se bons resultados utilizando como insumo as medições de um sistema localizado na Argélia.

Como apresentado, diversos são os estudos onde os algoritmos propostos no presente trabalho são utilizados para identificar falhas no funcionamento em equipamentos mecânicos, elétricos ou processos industriais. Sendo objeto de estudo a verificação de anomalia em máquinas utilizadas para produção de petróleo, fica evidente a razoabilidade da proposta ao sugerir a utilização de tais algoritmos.

É apresentada proposta única devido à necessidade de implementação de todos os algoritmos para eventualmente concluir-se qual a variação ótima, evitando a seleção de apenas um algoritmo caso as propostas fossem separadas. Custos financeiros atrelados não são alterados caso apenas uma ou todas as variações sejam implementadas, pois todas elas

encontram-se no escopo de atuação dos profissionais indicados, um fator positivo para manter proposta única de solução.

3.2 Conexão da proposta com os resultados esperados

A proposta de solução apresentada anteriormente foi selecionada para implementação dado ter sido a única. Apesar disso, ela propõe algumas variações importantes na sua aplicação que permitem chegar em uma solução ótima.

O desenvolvimento dos classificadores se deu pela plataforma *Jupyter Notebook* executada em ambiente dedicado disponibilizado pelo *Google Colab*, arquivo disponível em https://github.com/defante/mack-pos-inteligencia-artificial-tac, e a avaliação dos modelos, disponível no Apêndice B, teve como base as seguintes métricas:

- **Acurácia:** fração da quantidade de predições corretamente identificadas
- **Matriz de Confusão:** tabela que permite visualmente observar desempenho do classificador nas possíveis classes
- ***ROC Curve*:** curva de representação do desempenho de um classificador binário em diferentes limiares de classificação
- ***Recall*:** capacidade do modelo de identificar corretamente as observações de classe positiva dentre todas as verdadeiras observações de classe positiva
- ***Precision*:** capacidade do modelo de identificar corretamente as observações de classe positiva dentre todas as quais o classificador considerou como sendo de classe positiva
- ***F1-Score*:** métrica combinada do *Recall* e *Precision* que indica a capacidade do classificador em predizer as classes de observações

Os resultados encontrados para os classificadores estão muito satisfatórios e em linha com o objetivo proposto de gerar modelos capazes de identificar a anomalia estudada com acurácia razoável, onde os classificadores *Random Forest*, *Naive Bayes* e *Nearest Neighbors* obtiveram 98.97%, 92.29% e 99.86% de acurácia respectivamente ao serem testados contra 20% do volume de observações da base de dados e os outros 80% tendo sido utilizados para o treinamento. Apesar da sua performance, o *Random Forest* levou aproximadamente 50 minutos para ser treinado, o *Naive Bayes* apenas 5 segundos e o *Nearest Neighbors* aproximadamente 10 minutos. Se levarmos em consideração acurácia e tempo de treinamento, o *Naive Bayes* seria um ótimo candidato para aplicar retroalimentação durante sua execução com dados reais.

4 PLANEJAMENTO

Para devida implementação e utilização da proposta de forma com que seja adequadamente inserida no contexto da empresa, o presente capítulo tem como objetivo explicar todo planejamento e ações a serem tomadas para que se concretize.

4.1 Planos de ações

Nesse momento serão mencionadas todas as ações, em ordem cronológica de realização, e seus respectivos responsáveis para que a proposta de solução tenha sua viabilidade de implementação satisfeita.

Quadro 1 – Proposta de solução A: Combinação de classificadores do tipo *Random Forest, Naive Bayes* e *Forest Ensemble* para identificar anomalias do tipo instabilidade no fluxo

Objetivo: Estruturar implementação de modelos de classificadores para identificação de anomalia do tipo instabilidade no fluxo durante a produção de petróleo		
Ação Detalhada	**Prazo para Finalização/Implantação**	**Responsável (área/função)**
Análise dos dados já disponibilizados	5 dias	Especialista em Inteligência Artificial
Tratamento e limpeza	5 dias	Especialista em Inteligência Artificial
Criação dos classificadores (ciclo entre tratamento do dado, treinamento do modelo, avaliação e ajuste de parâmetros)	3 semanas	Especialista em Inteligência Artificial
Implementação dos modelos no sistema de monitoração em tempo real da empresa	3 semanas	Engenheiro de Software
Avaliação dos alertas e modelos	3 semanas	Engenheiro Responsável da Plataforma
Definição da variação ótima	5 dias	Especialista em Inteligência Artificial e Engenheiro Responsável da Plataforma

Fonte: Elaborado pelo autor

O engenheiro de software tem um papel chave nessa implementação, pois será responsável pela viabilização da utilização dos modelos em tempo real integrado com o sistema de monitoração da empresa, onde dados são recebidos constantemente dos sensores e os classificadores sejam acionados para devida identificação das observações.

Através desses passos será possível ao final do período a definição da variação ótima apresentada pela proposta de solução. Vale ressaltar que a fase de avaliação dos alertas e modelos pode se estender, a depender da frequência com que a anomalia ocorre e gerar eventos suficientes para uma avaliação razoável desse novo sistema.

Na plataforma o engenheiro responsável terá a incumbência de avaliar as variações implementadas dos modelos, conferindo se os alertas são procedentes ou se deixaram de alertar algum evento da anomalia e juntamente com o especialista em inteligência artificial definir a solução ótima, dado que o primeiro gerará insumos importantes para que o segundo consiga realizar as comparações necessárias entre a performance dos modelos e a combinação entre eles.

5 CONSIDERAÇÕES FINAIS

Este trabalho teve como objetivo a geração e avaliação de modelos de classificação binária para identificação de anomalia em plataformas de produção de petróleo *offshore* com intuito de contribuir com a disseminação e aplicação da inteligência artificial em problemas reais da indústria. Considerando a proposta de solução apresentada conclui-se que todos eles foram atingidos.

O *notebook* utilizado para construção dos modelos foi disponibilizado em repositório público mencionado anteriormente para que contribuições da comunidade possam ser recepcionadas e discutidas, melhorando as técnicas de tratamento de dados e treinamento dos algoritmos.

No que tange aos classificadores criados, o *Random Forest* e o *Nearest Neighbors* foram os quais obtiveram melhor acurácia e possivelmente performariam melhor com dados reais se comparados com os quais foram estudados neste trabalho. Apesar disso, o *Naive Bayes* também representa uma ótima solução quando pensamos em retroalimentação durante monitoração dos dados reais devido ao seu rápido tempo de treinamento e acurácia acima de 90%.

Com relação à continuidade existe uma possibilidade de evolução deste trabalho com a combinação de diversos classificadores binários, cada um especializado para identificar uma anomalia diferente. Outra possibilidade são as diversas variações nos tratamentos dos dados e alteração dos parâmetros de criação dos algoritmos.

Por fim, é possível concluir as diversas aplicabilidades da inteligência artificial para problemas relevantes ao nosso mundo das mais variadas naturezas, permitindo com que seus profissionais possam utilizá-la para que desafios modernos sejam superados trazendo evolução e benfeitoria à sociedade como um todo.

REFERÊNCIAS

AMIN, MD. TANJIN, et al. A Novel Data-Driven Methodology for Fault Detection and Dynamic Risk Assessment. Canadian Journal of Chemical Engineering, vol. 98, no. 11, pp. 2397–2416, 2020.

ANDREOLLI, I. Estabilidade Linear Aplicada ao Escoamento Multifásico de Petróleo. São Paulo. 2018.

BHATTACHARYA, S.; MISHRA, S. Applications of machine learning for facies and fracture prediction using Bayesian Network Theory and Random Forest: Case studies from the Appalachian basin, USA. Journal of Petroleum Science and Engineering, v. 170, p. 1005–1017, 2018.

HARROU, FOUZI, et al. Improved kNN-Based Monitoring Schemes for Detecting Faults in PV Systems. IEEE Journal of Photovoltaics, vol. 9, no. 3, pp. 811–821, 2019.

LI, HUIFANG, et al. Intelligent Fault Diagnosis for Large-Scale Rotating Machines Using Binarized Deep Neural Networks and Random Forests. IEEE Transactions on Automation Science and Engineering, vol. 19, no. 2, pp. 1109–1119, 2022.

MOULELOUED, YOUSSOUF, et al. A Developed Algorithm Inspired from the Classical KNN for Fault Detection and Diagnosis PV Systems. Journal of Control, Automation & Electrical Systems, vol. 34, no. 5, pp. 1013–1027, 2023.

VARGAS, R. E. V. et al. A realistic and public dataset with rare undesirable real events in oil wells. Journal of Petroleum Science and Engineering, v. 181, p. 106223, out. 2019. ISSN 09204105.

WEISS, GARY. Mining with rarity: A unifying framework. SIGKDD Explorations. 6. 7-19, 2004.

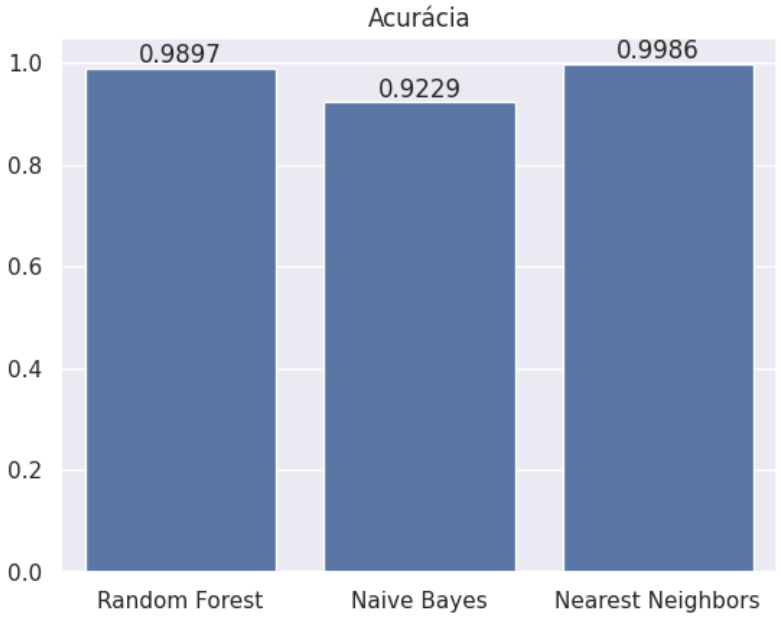
APÊNDICE A – Descrição da base de dados 3W

Os arquivos CSV que compõem a base de dados utilizada nesse estudo possuem a seguinte estrutura de colunas:

- *Timestamp*: momento em que a observação foi gerada
- P-PDG = Pressão no PDG [Pascal]
- P-TPT = Pressão no TPT [Pascal]
- T-TPT = Temperatura no TPT [Celsius]
- P-MON-CKP = Pressão a montante do PCK [Pascal]
- T-JUS-CKP = Temperatura a jusante do PCK [Celsius]
- P-JUS-CKGL = Pressão a jusante do GLCK [Pascal]
- T-JUS-CKGL = Temperatura a jusante do GLCK [Celsius]
- QGL = Fluxo de elevação de gás [Sm^3/s]
- *Class* = Classe da observação

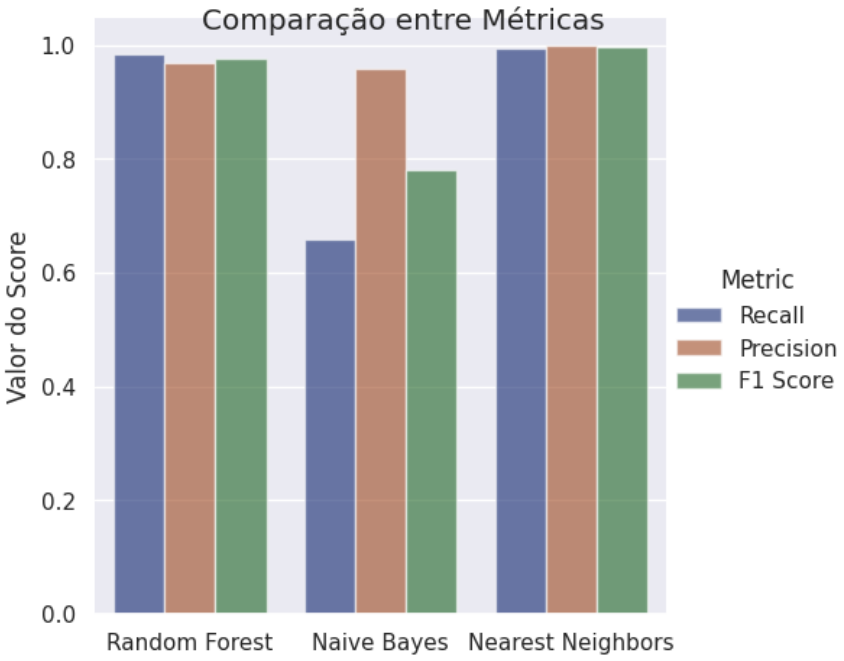
APÊNDICE B – Avaliação das métricas dos classificadores

Gráfico 3 – Comparação Acurácias



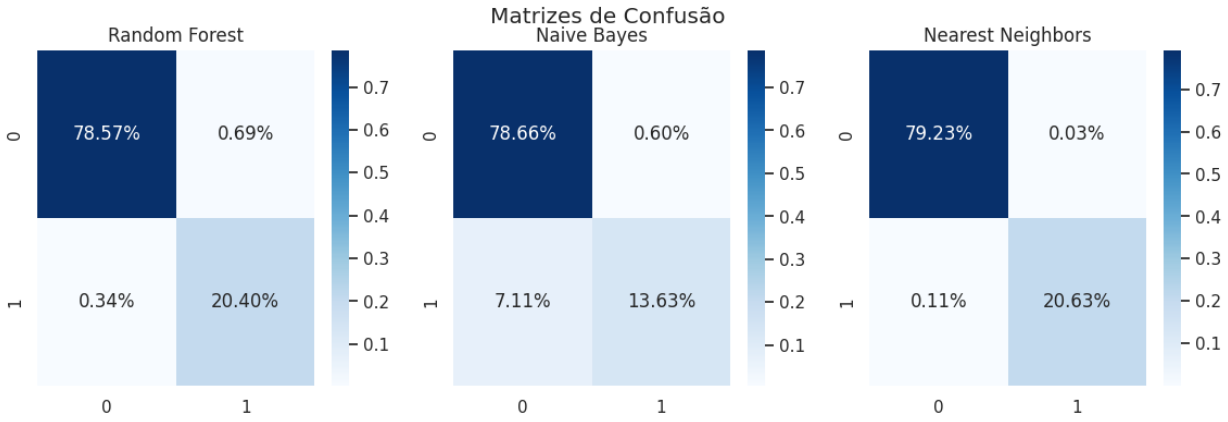
Fonte: Elaborada pelo autor

Gráfico 4 – Desempenho em outras métricas



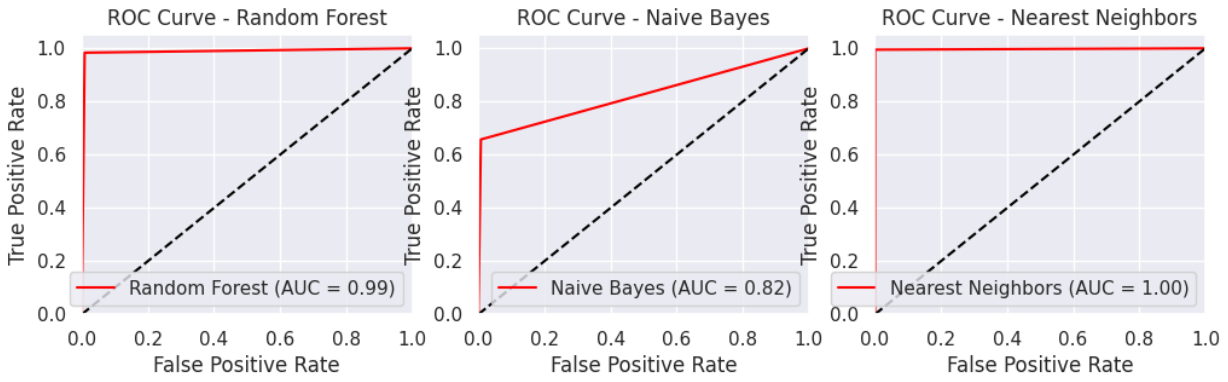
Fonte: Elaborada pelo autor

Figura 3 – Matrizes de Confusão



Fonte: Elaborada pelo autor

Gráfico 5 – *ROC Curves*



Fonte: Elaborada pelo autor