PONTIFÍCIA UNIVERSIDADE CATÓLICA
DO RIO DE JANEIRO

Deyvidy Luã de Oliveira Melo

Desenvolvimento de Método Para Detecção de Derivações Clandestinas Em Tubulações Baseado Em LIDAR.

Projeto de Graduação

Projeto de Graduação apresentado ao Departamento de Engenharia Mecânica da PUC-Rio

Orientador: Prof. Dr. Igor Braga de Paula
Coorientador: MSc. Igor Caetano Diniz

Rio de Janeiro
Novembro de 2024

Agradecimentos

Agradeço primeiramente a Deus, por me possibilitar chegar até aqui, após uma longa jornada de desafios e renúncias, com muitas conquistas e vitórias pelo trajeto, sem Ele nada disso seria possível, toda honra e glória ao Senhor Jesus Cristo.

Agradeço a minha família, que me apoiou no começo, meio e nesta reta final, com muito amor, paciência e carinho.

Agradeço à PUC por me apoiar até aqui, com minha bolsa, com uma estrutura de excelência e com todo o suporte desta renomada instituição.

Agradeço ao meu orientador, por aceitar este projeto com ânimo e dedicação.

Agradeço ao meu coorientador por aceitar este projeto e dedicar-se tanto em me ajudar a vencer os desafios envolvidos.

Agradeço aos professores que me instruíram nesta caminhada.

Agradeço a meus amigos e colegas que participaram ativamente desta conquista em tantos momentos de estudo conjunto e trabalho árduo.

RESUMO

Desenvolvimento de Método Para Detecção de Derivações Clandestinas Em Tubulações Baseado Em LIDAR.

O problema de perda de fluidos em linhas de transporte por meio de trincas, furos e derivações clandestinas é de interesse para diversas indústrias, sendo bastante crítico em sistemas de distribuição de água e no transporte de hidrocarbonetos. Nesse contexto, a detecção de furtos de combustíveis é de especial interesse para a indústria, pois tem impacto na segurança da população e na economia local. Sendo assim, é necessário o empenho em estudar e desenvolver possíveis soluções para mitigar esse problema. Nesse contexto, a metodologia proposta visa a detecção de vazamentos localizados. Para este estudo está sendo utilizado um sensor rotativo do tipo LIDAR para mapear a superfície interna de uma tubulação. O objetivo do estudo é avaliar alguns métodos de análise de dados como ferramenta de processamento das informações do sensor e detecção de anomalias. Assim, no presente projeto buscou-se detectar grandes anomalias através do mapeamento da geometria das paredes dos dutos. Para isso buscou-combinar o emprego da tecnologia LIDAR com técnicas de processamento digital dos dados e inteligência artificial. Os resultados obtidos se mostraram muito promissores e grandes anomalias puderam ser detectadas com elevado nível de acerto. Anomalias de geometria e dimensões conhecidas foram utilizadas como conjunto de treinamento para classificadores supervisionados, permitindo uma detecção gráfica a partir de parâmetros conhecidos dos algoritmos de machine learning utilizados.

Palavras chaves: LIDAR. Machine Learning. Detecção de Furtos. Derivações Clandestinas. Anomalias. Vazamentos localizados. Tubulações.

ABSTRACT

Development of a LIDAR-based method for the detection of clandestine derivations in pipelines

The problem of fluid loss in transport lines through cracks, holes, and clandestine derivations is of interest to various industries, being particularly critical in water distribution systems and hydrocarbon transport. In this context, the detection of fuel theft is of special interest to the industry, as it impacts public safety and the local economy. Therefore, it is necessary to study and develop possible solutions to mitigate this problem. In this context, the proposed methodology aims at detecting localized leaks. For this study, a rotary LIDAR sensor is being used to map the inner surface of a pipeline. The objective of the study is to evaluate some data analysis methods as a tool for processing sensor information and detecting anomalies. Thus, in the present project, it was sought to detect large anomalies through the mapping of the geometry of the duct walls. For this, it was sought to combine the use of LIDAR technology with digital data processing techniques and artificial intelligence. The results obtained were very promising, and large anomalies could be detected with a high level of accuracy. Anomalies of known geometry and dimensions were used as a training set for supervised classifiers, allowing graphical detection from known parameters of the machine learning algorithms used.

Key-words: LIDAR. Machine Learning. Theft detection. Clandestine derivations. Anomalies. Localized leaks. Pipes.

Sumário

Lista de Figuras

Lista de Tabelas

1 Introdução

De acordo com a ANP (2023), os dutos são estruturas fundamentais para o transporte de fluidos em geral, sendo utilizados em diversas aplicações, principalmente transporte de hidrocarbonetos. Em 2022, o Brasil possuía 589 dutos destinados à movimentação de petróleo, derivados de petróleo, gás natural e etanol, com um total de 20,2 mil km em tubulações. Deste total, 183 dutos, ou 14,4 mil km eram destinados ao transporte e 406, ou 5,8 mil km, destinados `transferência. Somente para a movimentação de gás natural, 113 dutos eram destinados a este combustível, enquanto isso, para os derivados de petróleo havia 416 tubulações disponíveis, totalizando 5,9 mil km, e para o petróleo bruto, havia 30 dutos, totalizando 2,3 mil km. Finalmente, para a movimentação de etanol havia 30 dutos, com uma extensão total de 450 km. [1].

Figura 1: INFRAESTRUTURA DE PRODUÇÃO E MOVIMENTAÇÃO DE PETRÓLEO E DERIVADOS - 2022. [1]

Na figura acima, é possível ver a infraestrutura nacional de produção e distribuição de petróleo e de derivados, implantada no Brasil em 2022.

O transporte por meio de dutos está sujeito a diversos problemas, como vazamentos, corrosão e furtos nas linhas de distribuição. Essas perdas de fluido podem causar perdas econômicas significativas, danos ambientais e acidentes. A corrosão é um dos fatores que pode levar a vazamentos, mas existem outros como furtos por meio de derivações clandestinas.

A detecção automática de anomalias em dutos se torna então importante e necessária para identificar problemas e agir na prevenção de eventos associados a perda de fluido transportado. Isso visa garantir o baixo custo e a alta confiabilidade no transporte por dutos, como exemplificado na Figura 2.

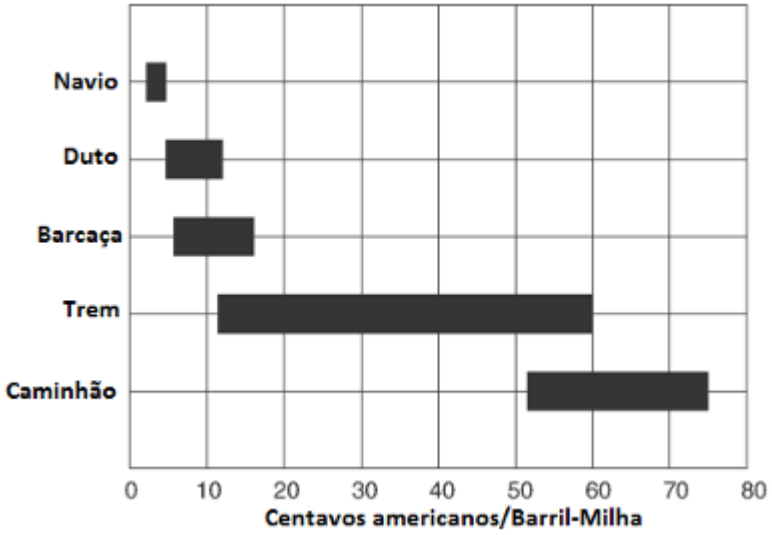

Figura 2: Relação de custo entre modais. Adaptado de (Liu 2003).

No contexto de monitoramento de dutos para detecção de anomalias nas tubulações é que o presente trabalho se insere. Diversas ferramentas para este propósito são disponíveis comercialmente e várias ainda se encontram em estágio de desenvolvimento. Neste trabalho foi avaliada uma metodologia alternativa para identificação de alterações nas paredes de um duto em condições de laboratório. Essa é uma das etapas iniciais para a prova de conceito da metodologia.

1.1 Objetivo

No presente projeto, o objetivo é detectar grandes anomalias através do mapeamento da geometria das paredes dos dutos. Para isso, pretende-se combinar o emprego da tecnologia LIDAR com técnicas de processamento digital dos dados e inteligência artificial através de técnicas de Machine Learning. A ideia é avaliar as dificuldades e capacidades da metodologia na detecção de anomalias em modelos de dutos com falhas conhecidas. Os objetivos específicos são listados a seguir:

(1) implementar um sistema de mapeamento bidimensional das tubulações utilizando sensor LIDAR;

(2) desenvolver algoritmos de processamento de dados que identifiquem anomalias nas superfícies inspecionadas;

Com o cumprimento dos objetivos espera-se obter uma plataforma para estudos mais elaborados de detecção de falhas específicas com LIDAR.

2 Revisão Bibliográfica

Nesta Seção faz-se uma revisão sucinta de alguns conceitos básicos dos métodos e dispositivos utilizados no âmbito deste trabalho.

2.1 LIDAR

Primeiramente é feita uma revisão breve acerca do dispositivo LIDAR, que foi o principal instrumento utilizado no desenvolvimento deste projeto.

2.1.1 Tecnologia LIDAR

Nesta seção serão abordados estudos e pesquisas relacionados ao uso do LIDAR, um tipo de sensor de distância que é baseado em laser (Light Detection and Ranging) como tecnologia promissora para a detecção de anomalias em tubulações. Chamado de maneira geral no contexto acadêmico de sensor LIDAR e em cenários militares como LADAR (Laser Detection And Ranging), o sensor é capaz de mapear em uma nuvem de pontos discreta o ambiente ao seu redor através da emissão de pulsos de luz e medição do tempo de retorno do sinal refletido (Bida; Padovezi Junior; Christoff, 2020).

A distância pode ser calculada de duas maneiras, sendo a primeira, calculando o tempo entre a emissão de um pulso de luz e o seu retorno após refletido pela superfície a ser investigada, sendo $c \approx 299.792.45\frac{m}{s}$ a velocidade da luz (Bida; Padovezi Junior; Christoff, 2020)

$$d = \frac{ct}{2} \tag{1}$$

Uma outra maneira, é a utilização da diferença de fase φ entre duas ondas, uma senoidal sendo a onda emitida e a outra, que também é uma senoidal, a onda refletida pela superfície investigada, sendo as duas ondas de mesma frequência f. Desta forma é possível determinar a distância entre o sensor LIDAR e a superfície alvo do pulso de luz (Bida; Padovezi Junior; Christoff, 2020).

$$d = \frac{c\varphi}{4\pi f} \quad (2)$$

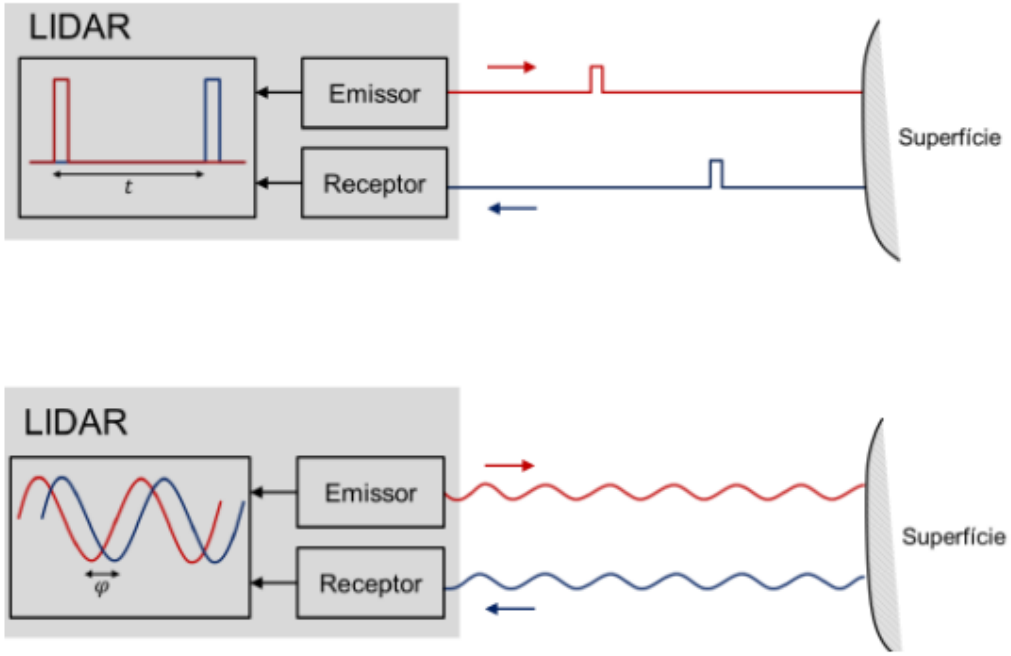

Figura 3: Esquema de funcionamento do LIDAR (FERREIRA, 2014)

Neste trabalho será utilizado um LIDAR com emissão e recepção de pulsos para se avaliar a capacidade de detecção de variações na seção radial de dutos que possam indicar a presença de vazamentos, obstruções ou mesmo derivações clandestinas

2.1.2 Aplicações do LIDAR em Diferentes Indústrias

Nesta subseção, são descritas brevemente algumas das aplicações do sensor LIDAR.

2.1.2.1 LIDAR em Estudos Hidrográficos

No trabalho desenvolvido por Soares e Galvíncio (2020), é demonstrado o potencial do LIDAR na caracterização de bacias hidrográficas, da bacia do Rio Beberibe, em Pernambuco. A caracterização detalhada da bacia foi realizada com base nos dados obtidos por um sensor LiDAR de alta resolução espacial (5m) e escala (1:5000), obtidos através

do programa PE3D (Pernambuco Tridimensional). A alta resolução espacial dos dados obtidos com o LiDAR (em escala geográfica) permitiu uma análise bem detalhada e precisa de toda a superfície da bacia, contribuindo para um melhor entendimento do comportamento hidrológico e dos processos hídricos naquela área, de forma que o uso do LiDAR naquele estudo se mostrou interessante para auxiliar possíveis tomadas de decisão na gestão dos recursos hídricos da região.

2.1.2.2 LIDAR em Mapeamento Florestal

O artigo de Giongo (2010) oferece uma revisão bastante abrangente sobre os princípios e aplicações do uso do sensor LIDAR na área florestal, onde nesse trabalho em questão é destacada a importância das informações sobre topografia e cobertura florestal para gestão dos recursos florestais e naturais, ressaltando a dificuldade em se obter dados precisos sobre altura das árvores e densidade florestal com técnicas convencionais. Assim, o uso do LIDAR aparece como uma excelente e bastante promissora solução para superar essas dificuldades.

Giongo (2010) ainda destacou que houve um aumento significativo nas pesquisas sobre aplicações do LIDAR nesta área, impulsionado pelas vantagens do sensor na captura de informações tridimensionais da superfície terrestre, especialmente em áreas complexas e de difícil acesso.

No caso de dutos, o acesso também é difícil e sensores do tipo LIDAR podem vir a ser uma alternativa promissora no mapeamento da estrutura interna das tubulações.

2.1.2.3 LIDAR em Monitoramento de Gasodutos Após Desastres Naturais

O estudo de Gong (2016), fruto de uma colaboração entre o Centro de Infraestrutura e Transporte Avançado da Rutgers University e o Gas Technology Institute, teve como objetivos investigar a aplicação de um sistema móvel híbrido que une o uso do LIDAR com o infravermelho para

monitoramento aéreo de gasodutos, com foco na detecção de ameaças e na avaliação de riscos após desastres naturais. O trabalho se debruça em cima da preocupação com a integridade da vasta rede de gasodutos nos Estados Unidos, especialmente diante da intensificação de eventos climáticos extremos.

A área de estudo de Ortley Beach em New Jersey, que foi afetada pelo furacão Sandy, foi utilizada como base para o estudo. Naquele trabalho, um mapa da localização do sistema de distribuição de gás antes da passagem do furacão (tubulações principais, linhas de serviço e risers de medidores) foi comparado com o mapa medido após o desastre. Assim, os pesquisadores avaliaram o movimento do solo e as mudanças no nível da água, de forma que os movimentos dos edifícios foram obtidos a partir dos dados gerados pelo LIDAR. Com esses dados eles traçaram mapas de deformação e tensão dos tubos e classificaram os que tinham mais probabilidade de dano. A análise envolveu o uso de ferramentas de elementos finitos para o cálculo dos esforções nos dutos permitindo assim estimar uma probabilidade de danos.

2.2 Processamento de Dados do LIDAR

De forma geral, o LIDAR rotativo gera uma nuvem de pontos em um processo de varredura que depende da frequência de amostragem de dados e da velocidade de rotação do sensor. Isso gera uma nuvem de pontos contendo a distância do objeto no meio para o sensor LIDAR e a posição angular do sensor. Essas informações, por sua vez, correspondem a superfície a ser medida. Essas nuvens de pontos podem ser melhor analisadas com técnicas de Machine Learning (SANTOS, 2023).

2.2.1 Algoritmos de Detecção

Nesta seção são descritos alguns dos algoritmos de aprendizado de máquina empregados no trabalho para a classificação dos dutos com e sem a presença de anomalias.

2.2.1.1 Random Forest

O Random Forest, proposto originalmente por Breiman (2001), constitui-se como um método de aprendizado de máquina supervisionado, caracterizado como um algoritmo de classificação e regressão baseado em ensemble de árvores de decisão.

Na concepção metodológica do Random Forest, o algoritmo CART (Classification and Regression Trees) emerge como fundamento essencial para construção das estruturas de classificação. O processo de desenvolvimento das árvores caracteriza-se pela expansão completa, sem procedimentos de poda, mantendo a estrutura original em sua máxima profundidade informacional [15].

O algoritmo fundamenta-se em três componentes metodológicos principais. O Bootstrap Aggregating (Bagging), conforme descrito por Rodriguez-Galiano (2012), onde o bagging permite a criação de múltiplos subconjuntos de treinamento através de reamostragem aleatória com reposição, garantindo diversidade entre as árvores geradas. Assim, geram-se reamostras bootstrap, que são amostras sorteadas do conjunto original, de mesmo tamanho e via sorteio simples com reposição, onde cada subconjunto amostrado é utilizado para a construção de um novo classificador (Breiman, 1996). A classificação final então é realizada por um sistema de votação

A ideia da distribuição via bootstrap é que quando só conseguimos obter uma amostra da população (caso típico), usamos reamostragem bootstrap como um substituto para outras possíveis amostras da população (Hesterberg, 2003).

Outro componente é a Seleção Aleatória de Características (Strobl, 2007), fazendo com que em cada nó de divisão, apenas um subconjunto randômico de características é considerado, reduzindo correlações entre árvores e melhorando a capacidade preditiva.

Por último, na etapa de classificação, ocorre uma agregação dos resultados individuais das árvores através de uma votação majoritária, onde

a classe mais frequente determina o resultado final (CRIMINISI; SHOTTON; KIPMAN, 2012).

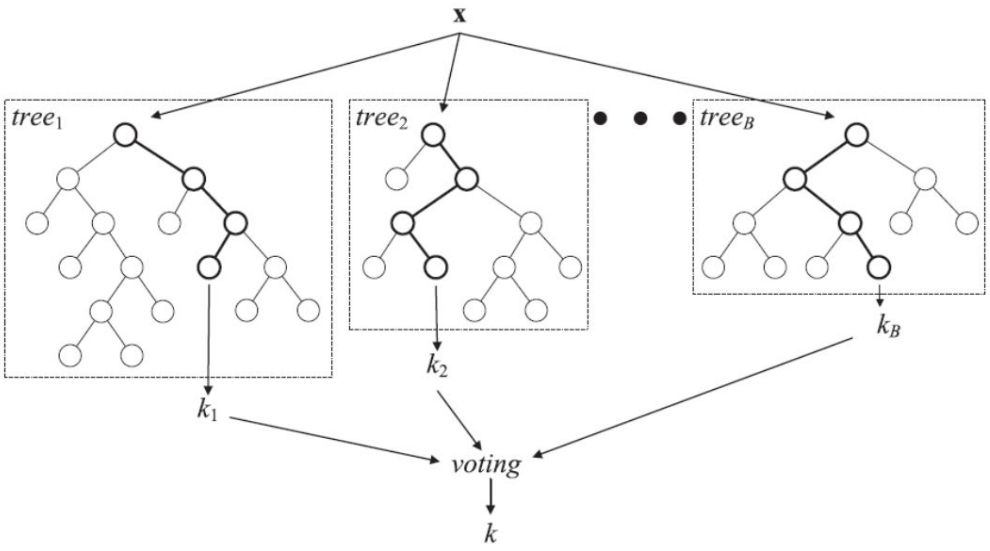

Figura 4: RF Baggin Scheme - Adaptado de Hastie, Tibshirani, Friedman (2009, Cap.8)

Considerando as contribuições fundamentais de Breiman [4] e [5] para modelagem estatística, aqui é demonstrada a definição formal do conjunto de treinamento e seus componentes, onde, seja L um conjunto de dados de treinamento representado pelo conjunto {(xi,yi), $,i$ = 1,2,...,N}, N representa o número total de exemplos no conjunto, xi corresponde ao vetor de atributos de cada exemplo e yi representa a classe associada ao i-ésimo exemplo, pertencente ao conjunto {1,2,...,K}. O preditor de classe para um elemento x, derivado do conjunto de treinamento L, será denominado $\psi(\bullet,\mathrm{L})$. Propõe-se uma estratégia de reamostragem bootstrap, na qual são geradas V amostras independentes $\{\mathrm{L}^{(v)}\}$, onde v = 1,2,...,V. Cada amostra contém N elementos independentes que foram extraídos da distribuição original do conjunto L. O objetivo metodológico consiste em induzir V preditores distintos a partir dessas amostras $\{\mathrm{L}^{(v)}\}$, aplicar um método de agregação para aprimorar a capacidade preditiva e então representar a sequência de preditores como $\psi(\bullet,\mathrm{L}^{(v)})$.

Para um conjunto de classes indexadas por k = 1, 2, ..., K, propõe-se então um método de combinação dos preditores individuais $\psi(\bullet,\mathrm{L}^{(v)})$ através

de mecanismo de votação majoritária, contabilizando assim o número de votos para cada classe e a classe com maior número de votos será selecionada como resultado final e atribuída a x.

Assim, seja N_k o total de votos recebidos pela classe k

$$N_k = |\{v \in \{1 \dots V\}. \psi\left(x, L^{(v)}\right) = k\}| \tag{3}$$

O classificador agregado será definido como:

$$\psi_A = a_{\underset{k}{rgmax}\, N_k} \tag{4}$$

Onde o subscrito A representa o processo de agregação [3].

O método de Obtenção de $\{\mathrm{L}^{(v)}\}, v$ = 1,2,...,V será realizado com a utilização de reamostragem bootstrap de L através de sorteio com repetição, com cada uma de tamanho N, e com isso, este procedimento permite criar um modelo ensemble com maior capacidade preditiva e robustez.

Na geração de cada amostra bootstrap, são utilizados aproximadamente 63% dos exemplos originais para construir a árvore.

$$\begin{aligned} \Pr(Observação\ i \in amostra\ bootstrap\ b) &= 1 - \left(1 - \frac{1}{N}\right)^N \\ &\approx 1 - e^{-1} = 0.632 \end{aligned} \tag{5}$$

Logo, alguns exemplos são deixados de fora e não serão utilizados na indução do classificador. Esse conjunto de exemplos é chamado de Out-Of-Bag (OOB) e é usado no teste do classificador construído para estimar o erro em cada árvore da floresta. Este processo evita então qualquer necessidade de realização de cross-validation ou de testes separados com outro conjunto de dados [4].

O erro *(OOB)* de cada classificador $\psi(\bullet, \mathrm{L}^{(v)})$ é definido então como o percentual do conjunto de teste (constituído por $\mathrm{L} \backslash \mathrm{L}^{(v)}$) erroneamente classificado.

2.2.1.2 KNN

O algoritmo K-Nearest Neighbors (KNN), ou K-Vizinhos Mais Próximos, é um método supervisionado de Machine Learning, amplamente utilizado para tarefas de classificação e regressão [6]. Sua popularidade se deve à sua simplicidade conceitual, facilidade de implementação e capacidade de lidar com dados não lineares [7]. Apesar de ser um dos algoritmos mais antigos em aprendizado de máquina, o KNN continua sendo uma ferramenta poderosa e relevante em diversas aplicações, como reconhecimento de padrões, mineração de dados e análise preditiva [8].

3 Fundamentos do KNN

O princípio fundamental do KNN é a ideia de que pontos de dados semelhantes tendem a estar próximos uns dos outros no espaço de características [9]. Ou seja, o algoritmo assume que a classe ou valor de um ponto de dado desconhecido pode ser inferida a partir da classe ou valor de seus vizinhos mais próximos.

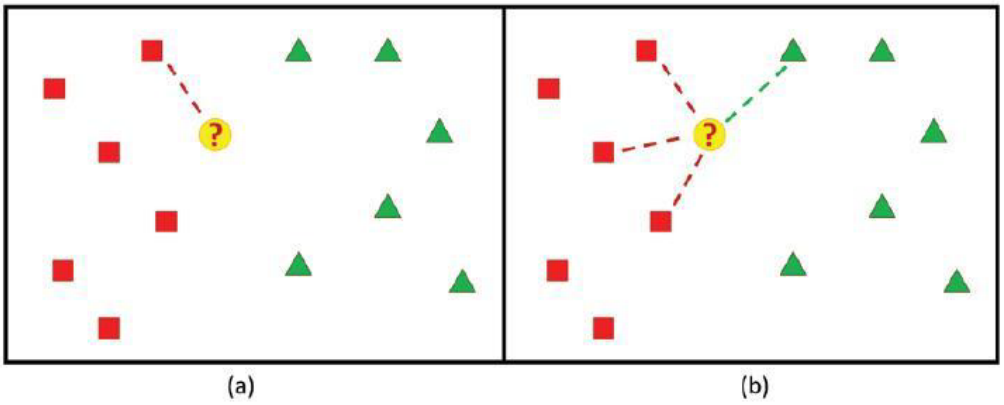

Figura 5: (a) The 1-NN decision rule. (b) the KNN (IMANDOUST, 2013)

Em problemas de classificação, o KNN atribui a um novo ponto de dado a classe mais frequente entre seus k vizinhos mais próximos. A Figura acima ilustra esse processo de classificação com k=1 (Figura 5-a) e com k=4 (Figura 5-b), onde pode-se observar que entre as amostras mais próximas, temos três amostras que são Quadrados Vermelhos e uma amostra que é

Retângulo Verde, então a nova amostra desconhecida, de Círculo Amarelo é na verdade da classe do Quadrado Vermelho (Figura 5-b). O parâmetro k, que representa o número de vizinhos mais próximos considerados, é crucial para o desempenho do KNN. Um valor pequeno de k pode levar a overfitting (superajuste), onde o modelo se ajusta muito aos dados de treinamento e não generaliza bem para novos dados. Por outro lado, um valor grande de k pode levar a underfitting (subajuste), onde o modelo é muito simplista e não captura a complexidade dos dados [11]. Neste trabalho serão utilizados k=5 para o modelo sem SMOTE e k=10 para o modelo com SMOTE, e esses são valores adequados e padronizados para com uso e sem uso de oversampling [13].

Em problemas de regressão, o KNN estima o valor de um novo ponto de dado calculando a média (ou mediana) dos valores de seus k vizinhos mais próximos, primeiramente calculando a distância entre o novo ponto de dado e todos os pontos de dados no conjunto de treinamento. As métricas de distância mais comuns incluem a distância Euclidiana, distância de Minkowski, City Block e Chebychev. A escolha da métrica de distância pode influenciar o desempenho do algoritmo. Neste modelo é utilizada a distância euclidiana Eq. (6), pois é a padrão da biblioteca Scikit-Learn e uma das mais assertivas (CHOMBOON, 2015).

$$d_{st} = \sqrt[p]{\sum_{j=1}^{n} |x_{sj} - y_{tj}|^p} \tag{6}$$

A distância de Minkowski é um método para cálculo de distância baseado no espaço Euclidiano e definido pela Eq, (4). A partir dele são obtidas algumas das métricas mais comuns citadas acima ao variar o valor de p, como explicitado abaixo:

Para p = 1, temos a distância City Block [13].

$$d_{st} = \sum_{j=1}^{n} |x_{sj} - y_{tj}| \tag{7}$$

Para p = 2, temos a distância Euclidiana, explicitada abaixo de duas formas distintas [13].

$$d_{st} = \sqrt{\sum_{j=1}^{n} |x_{sj} - y_{tj}|^2} \quad (8)$$

ou

$$d_{st} = (x_s - y_t)(x_s - y_t)' \quad (9)$$

E para p = ∞, temos a distância Chebychev [13].

$$d_{st} = max_j\{|x_{sj} - y_{tj}|\} \quad (10)$$

Com as distâncias calculadas, o KNN seleciona os k pontos de dados com as menores distâncias em relação ao novo ponto de dado e então atribui ao novo ponto de dado a classe mais frequente entre seus k vizinhos mais próximos. Em caso de empate, podem ser utilizadas estratégias de desempate, como ponderação por distância, e então é calculada a média (ou mediana) dos valores dos k vizinhos mais próximos para estimar o valor do novo ponto de dado. [10].

O K-Nearest Neighbors apresenta vantagens e desvantagens que tornam sua aplicação específica a certos contextos. Entre suas vantagens, destacam-se sua simplicidade conceitual e facilidade de implementação, que o tornam acessível mesmo para usuários iniciantes, além de sua capacidade de lidar com dados não lineares. O KNN não exige um treinamento explícito, pois o modelo é construído diretamente com base nos dados de treinamento. No geral, é versátil, sendo aplicável tanto para tarefas de classificação quanto de regressão. Por outro lado, o algoritmo apresenta algumas desvantagens, pois é sensível à escolha do valor de k e seu custo computacional pode ser elevado em grandes conjuntos de dados devido à necessidade de calcular a distância entre cada novo ponto e todos os pontos do conjunto de treinamento. Além disso, o KNN é vulnerável à presença de ruído e outliers (valores discrepantes) e pode enfrentar dificuldades em

espaços de alta dimensionalidade devido à "maldição da dimensionalidade", que reduz a informatividade das distâncias entre os pontos [12].

Ao analisar todo o potencial do KNN, é notório que este é um algoritmo de aprendizado de Machine Learning intuitivo e com possibilidade de aplicações em diversas áreas, apesar de suas limitações, como a sensibilidade ao valor de k e a complexidade computacional, o KNN continua sendo uma ferramenta valiosa para análise de dados e modelagem preditiva. A escolha do KNN como método de aprendizado para este trabalho é baseada nas características do problema, nos requisitos de desempenho e na disponibilidade de recursos computacionais.

3.1.1.1 Logistic Regression

A Regressão Logística é um método estatístico amplamente utilizado para modelar a probabilidade de ocorrência de um evento binário (dicotômico) [14]. Embora o nome sugira uma técnica de regressão, a Regressão Logística é, na verdade, um algoritmo de classificação que prevê a probabilidade de um dado pertencer a uma determinada classe [15]. Suas aplicações são vastas, abrangendo áreas como medicina, engenharia, matemática aplicada, finanças, marketing e ciências sociais, sempre que o objetivo é analisar a relação entre variáveis independentes (preditoras) e uma variável dependente binária (resultado) [16]. A Regressão Logística prevê a probabilidade de um evento ocorrer e para isso utiliza a função logística (ou sigmoide), que transforma uma combinação linear das variáveis preditoras em um valor entre 0 e 1, representando a probabilidade [17].

Segundo Hosmer e Lemeshow (1989), muitas distribuições têm sido propostas para serem utilizadas em análises de uma variável em resposta binária ao longo dos anos para atender a desafios que dependam de técnicas de classificação.

Uma das principais razões para a escolha da distribuição logística é que, do ponto de vista matemático, a função é de utilização fácil e flexível.

Assim, tomando Y como uma variável resposta que pode assumir só dois valores, representados por (Y = 1) como sucesso e (Y = O) como fracasso, o valor esperado de Y é dado por

$$E(Y) = P(Y = 1) = \pi \tag{11}$$

A Eq. (11) denota a probabilidade de ocorrência de um evento (Y=1).

A distribuição condicional de uma variável resposta Y segue uma binomial com probabilidade dada pela média condicional $\pi(x) = E(Y|x)$. Desta forma, a probabilidade de sucesso de uma variável dependente Y, sendo o vetor das variáveis independentes $x = \{x_1, x_2 \dots x_p\}$, é representado por $P(Y = 1|x) = \pi(x)$ logo, a probabilidade de fracasso é dada por $P(Y = 0|x) = 1 - \pi(x)$.

O modelo de logistic Regression mais tradicional é dado pela Eq. (12), sendo g(x) a transformação logit, uma função linear que varia de -∞ a +∞ e é dada pela Eq. (13), onde β são os parâmetros do modelo, e esses parâmetros serão estimados pelo método da máxima verossimilhança. Para aprofundamento neste procedimento de estimação dos parâmetros do modelo logístico, àqueles com maior interesse é recomendado ao leitor Carballo (2002) e HOSMER JR, D. W.; LEMESHOW, S.; STURDIVANT, R. X. (2013).

$$\pi(x) = \frac{e^{g(x)}}{1 + e^{g(x)}} \tag{12}$$

$$g(x) = \beta_0 + \beta_1 x_1 + \cdots + \sum_{l=1}^{k_j - 1} \beta_{jl} x_{jl} + \beta_p x_p \tag{13}$$

A relação entre uma única variável x e a função logit g(x), graficamente possui um comportamento em forma de “S”, característico do modelo logístico Lemeshow (1989).

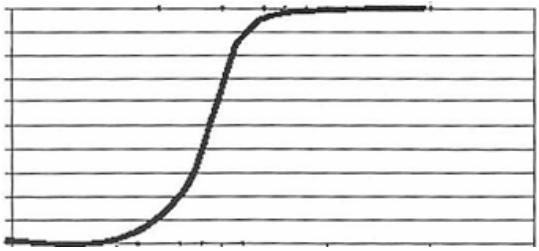

Figura 6: Ilustração da função Logística em Forma de "S"

A forma mais utilizada de interpretar os coeficientes do modelo logístico é utilizando a razão de chances (*odds ratio,* em inglês). Em modelos com variável resposta e apenas uma covariável binária, a chance de resposta está presente entre os indivíduos que apresentam x=1 e é definida como

$$\frac{\pi(1)}{1-\pi(1)} \tag{14}$$

Assim, a razão de chances ψ é demonstrada na Eq. (15) e consequentemente na Eq. (16).

$$\psi = \frac{\frac{\pi(1)}{1-\pi(1)}}{\frac{\pi(0)}{1-\pi(0)}} \tag{15}$$

$$\psi = \frac{e^{\beta_0+\beta_1}}{e^{\beta_0}} = e^{\beta_1} \tag{16}$$

3.2 Principais Técnicas De Detecção De Anomalias Em Dutos

A inspeção de dutos é crucial para a segurança e eficiência no transporte de fluidos como petróleo e gás. Sendo assim, para garantir a integridade da tubulação e prevenir vazamentos, diversas técnicas são empregadas, cada uma com suas características e vantagens específicas.

3.2.1 Inspeção Humana

Essa técnica é a mais simplista e tradicional que se mantém relevante, especialmente para dutos aéreos ou em áreas acessíveis e pode ser realizada através de um indivíduo por suas próprias percepções, envolvendo a verificação periódica da faixa de dutos por diferentes meios, como a pé, de carro ou por via aérea. Essa técnica permite a observação de indícios visuais de vazamentos, como alterações na coloração do solo, formação de bolhas em corpos d'água e distorções em imagens térmicas. Profissionais treinados podem identificar o odor característico do gás mesmo em concentrações baixas, além de detectar o som do vazamento a uma curta distância. No entanto, a inspeção humana apresenta desafios em termos de logística, cobertura da área e confiabilidade, tornando sua aplicação menos frequente. Apesar das limitações continua sendo uma ferramenta importante, especialmente quando combinada com outras técnicas. (GEREMIA, 2012)

3.2.2 Ultrassom

O ensaio por ultrassom é um dos métodos mais utilizados na inspeção de dutos, permitindo a detecção de descontinuidades internas e a medição da espessura da parede do duto. O método baseia-se na emissão de ondas ultrassônicas que penetram no material e retornam ao sensor, permitindo a análise do tempo de retorno e da intensidade do sinal para identificar perdas de material e defeitos. O ultrassom pode ser aplicado tanto internamente, por meio de PIGs ultrassônicos, quanto externamente, com sensores em contato com a parede do duto. A técnica é versátil, de alto custo, com elevado volume de dados para processamento mas que permite a detecção de diferentes tipos de defeitos, como corrosão, trincas e laminação, auxiliando na avaliação da integridade do duto e na tomada de decisões de manutenção e reparo. (GEREMIA, 2012)

3.2.3 Modelagem Matemática

A modelagem matemática é uma técnica que utiliza dados de pressão, vazão e temperatura do fluido para criar modelos computacionais que simulam o comportamento do fluido no interior do duto. Esses modelos permitem a detecção de vazamentos e anomalias no padrão de fluxo, inclusive vazamentos pequenos e imperceptíveis por outros métodos. A modelagem matemática também auxilia na previsão de falhas e na otimização das operações de transporte, identificando pontos de risco e prevendo a vida útil do duto. No entanto, a técnica requer dados precisos e um modelo bem calibrado para garantir a confiabilidade dos resultados. (COLOMBAROLI, 2009)

3.2.4 Correntes Parasitas

O ensaio por correntes parasitas é um método não destrutivo que utiliza um campo magnético gerado por uma sonda ou bobina para induzir correntes elétricas na peça ensaiada. A presença de descontinuidades ou variações nas propriedades do material altera o fluxo das correntes parasitas, permitindo a detecção de defeitos como trincas, corrosão e variações na espessura. A técnica é rápida, limpa e de baixo custo, sendo utilizada para inspeção de diferentes componentes, como tubos, barras, parafusos e soldas. (GEREMIA, 2012)

3.2.5 Tomografia Magnética

A tomografia magnética é um método não destrutivo que permite a inspeção de dutos metálicos ferromagnéticos, como os utilizados no transporte de petróleo e gás. A técnica baseia-se na aplicação de um campo magnético externo e na medição da resposta do material, permitindo a detecção de descontinuidades, variações de espessura e outros defeitos. A tomografia magnética é capaz de identificar a posição e a profundidade dos defeitos, fornecendo informações importantes para a avaliação da

integridade do duto e para o planejamento de manutenções e reparos. (OLIVEIRA; FARIAS; CABRAL, 2016)

3.2.6 Vantagens e Limitações do LIDAR em Comparação com Outras Tecnologias de Detecção em Pipelines

Já foi exposto nos tópicos iniciais deste trabalho como a detecção de vazamentos em pipelines é crucial para garantir a segurança operacional, a proteção ambiental e a eficiência econômica. Tendo isto claro, é importante realizar uma análise comparativa do LIDAR com suas vantagens e desvantagens frente as diferentes tecnologias de detecção de vazamentos disponíveis no mercado, das quais as principais estão descritas nos subtópicos imediatamente anteriores a este.

De acordo com Maia (2024), o LIDAR como um método de detecção e coleta de dados que pode ser baseado em IoT (Internet of Things), pode ser uma excelente escolha para um sistema automático ou semi-automático de inspeção na pipeline. E como a tecnologia é tida como uma aplicação baseada em hardware que pode ser utilizada de forma interna ou externa na tubulação, com o uso de PIG de maneira interna ou de forma externa pelo meio aéreo com o uso de aviões/drones ou terrestre com o uso de veículos

Porém quando comparado com algumas técnicas já existentes nos deparamos com uma limitação operacional importante quando se trata de uso interno em PIG com captação em tempo real, que é o fato de não poder ser utilizado em longas distâncias, uma vez que a aquisição dos dados perderia a conexão com o receptor (MAIA, 2024). No entanto, o volume de dados da nuvem de pontos é relativamente mais baixo em comparação com sensores ultrassônicos. Além disso, a análise da topologia da superfície pode ser rapidamente feita com o uso de ferramentas de IA. Logo, a tecnologia pode permitir o processamento das informações in-loco e em tempo real fazendo com que o volume de dados armazenados ou transmitidos seja bastante reduzido.

3.3 PIG

Nesta seção faz-se uma breve revisão dos dispositivos de inspeção de dutos (PIG - pipeline inspection gauge).

3.3.1 Dispositivo PIG

O "pig" é um dispositivo utilizado para a inspeção interna de dutos e tubulações. Ele é projetado para percorrer o interior dos dutos, realizando a avaliação da integridade da tubulação, identificando anomalias e coletando dados relevantes. O funcionamento do "pig" varia de acordo com o tipo de inspeção desejada, mas, em termos gerais, o processo envolve o dispositivo sendo inserido no duto, onde ele se move ao longo do pipeline, utilizando sensores para coletar informações sobre o estado da estrutura interna do duto (CAMERINI, 2018 e GEREMIA, 2012).

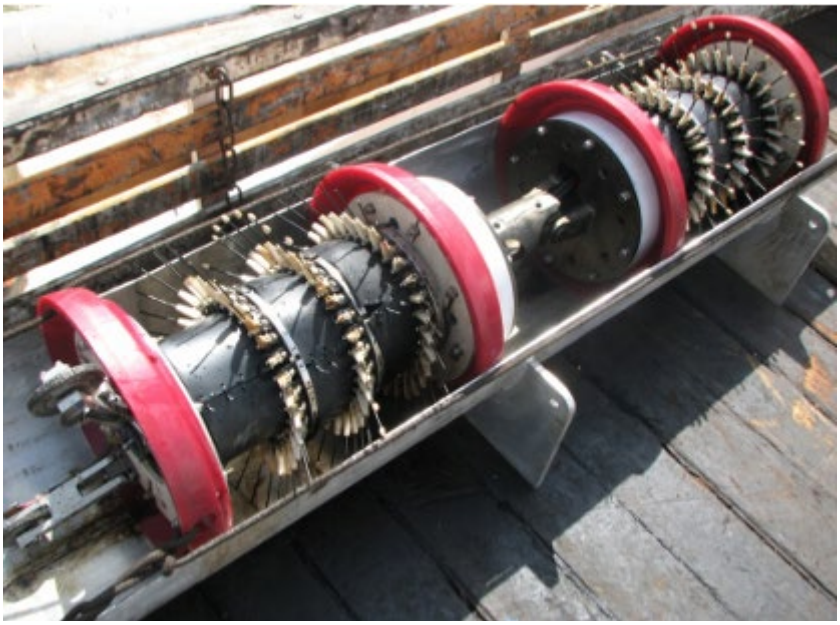

Figura 7:Pig de 16 polegadas usado pela Petrobras (Souza, 2010)

Dependendo do modelo e da tecnologia empregada, o "pig" pode ser equipado com sensores de diversos tipos, como ultrassom, magnetômetro, sensores de deformação ou até mesmo câmeras e sensores como o LIDAR para realizar a inspeção visual. Com o uso do "pig" é possível detectar diversos tipos de falhas, como corrosão, fissuras, vazamentos ou outras anomalias estruturais, fornecendo dados cruciais para a manutenção

preventiva e corretiva dos sistemas de tubulação, minimizando o risco de falhas catastróficas e melhorando a segurança operacional das tubulações (GEREMIA, 2012).

Segundo GEREMIA (2012), esse tipo de inspeção é considerado um método eficaz, pois permite a coleta de dados detalhados e na maioria das vezes, sem a necessidade de interrupções no funcionamento do sistema, o que representa uma grande vantagem em termos de continuidade operacional e redução de custos.

3.3.2 Funcionamento do PIG

O funcionamento do "pig" varia de acordo com o tipo de inspeção desejado, em uma visão geral do processo de inspeção, este, começa com a inserção do "pig" na tubulação, que é feito através de uma estação de lançamento localizada normalmente em uma extremidade do duto. Dependendo do tipo de "pig", ele pode ser propelido pelo fluxo do fluido no duto, como no caso dos "pigs" in-line, ou ser impulsionado mecanicamente, (GEREMIA, 2012).

Após a inserção, o "pig" se move ao longo do duto. Para os "pigs" in-line, o movimento é facilitado pelo fluxo do fluido, enquanto os "pigs" batch ou gel podem ser empurrados por uma pressão adicional ou por motores internos, proporcionando maior controle sobre sua movimentação (GEREMIA, 2012).

As informações coletadas durante a inspeção, como a espessura das paredes do duto, a presença de corrosão e a localização de derivações clandestinas, são registradas pelo "pig" e armazenadas para posterior análise. Por fim, ao atingir a extremidade do duto ou uma estação de recebimento, o "pig" é retirado do sistema, e os dados coletados são transferidos para sistemas de análise e interpretação, onde são avaliados para identificar a condição do duto e a necessidade de manutenção ou reparos (CAMERINI, 2004 e GEREMIA, 2012).

3.3.3 Aplicações

Existem diferentes tipos de pigs, como os pigs de limpeza e os pigs de inspeção. Os pigs de limpeza são utilizados para remover resíduos e detritos acumulados dentro do duto, enquanto os pigs de inspeção têm a função de avaliar a condição estrutural do duto. Esses últimos são frequentemente equipados com uma variedade de sensores capazes de detectar anomalias no interior dos dutos, como sensores de pressão, temperatura e vibração. Esses sensores possibilitam a identificação precoce de falhas, como vazamentos, corrosão, deformações ou fraturas, garantindo a integridade e segurança da infraestrutura (CAMERINI, 2004).

As aplicações mais comuns dos pigs incluem a detecção de vazamentos, que pode ser feita por meio da medição de variações na pressão ou temperatura do fluido que circula pelo duto. A corrosão é outra falha frequentemente detectada, com os pigs identificando alterações na espessura das paredes do duto. Além disso, deformações e fraturas podem ser detectadas com precisão, devido à capacidade do pig de monitorar as alterações estruturais do duto, como descontinuidades e variações no formato original. Esses dispositivos oferecem uma série de vantagens em relação a outras técnicas de inspeção. São considerados não intrusivos, pois não requerem a abertura do duto, o que reduz significativamente o risco de danos à estrutura do sistema. Além disso, sua operação é rápida e eficiente, permitindo a inspeção de longos trechos de duto em um curto espaço de tempo. A precisão e confiabilidade das medições realizadas pelos pigs também são fatores determinantes para a escolha dessa tecnologia (GEREMIA, 2012).

Contudo, o uso de pigs também apresenta desvantagens, como o custo elevado, que pode variar de acordo com o tipo de pig e os sensores empregados. Além disso, a instalação e operação do pig possui um nível de complexidade alto, exigindo treinamento especializado para garantir o bom funcionamento do dispositivo.

3.3.4 Utilização do PIG com LIDAR

O uso de "pigs" com tecnologia LIDAR é especialmente valioso, pois essa abordagem fornece dados bi e tridimensionais detalhados, podendo fornecer informações acerca da condição interna dos dutos. Esses dados são importantes para a identificação precoce de problemas e a programação de manutenção preventiva para a garantia da integridade e segurança do sistema de tubulações.

Um exemplo de avanço nesta área pode ser encontrado em pesquisas como o artigo “A Simultaneous Pipe-Attribute and PIG-Pose Estimation (SPPE) Using 3-D Point Cloud in Compressible Gas Pipelines”, publicado no periódico Sensors (NGUYEN; PARK; JEONG, 2023), onde os autores abordam os desenvolvimentos recentes em tecnologias de inspeção de dutos, incluindo o uso de "pigs" com sensores Lidar. Esses avanços destacam a importância crescente de métodos não intrusivos para a inspeção de dutos, visando otimizar a detecção de anomalias e aumentar a eficiência operacional eficiência operacional na indústria de dutos, enfatizando o uso do LIDAR, na otimização da detecção de falhas e na e destacando suas vantagens em relação às técnicas tradicionais.

4 Metodologia

Para o desenvolvimento do presente projeto é utilizado um sensor LIDAR no centro de um duto de 300mm, em um primeiro momento em posição estática e posteriormente com variação de posição e ângulo ao longo deste duto para simular o cenário real, onde então furos de diâmetros conhecidos que foram introduzidos para averiguar a sensibilidade do sensor e como os dados são captados.

Em um segundo momento criado um algoritmo em linguagem Python com uso dos algoritmos Random Forest, KNN e Logistic Regression para classificar os resultados obtidos em tubulações com diferentes anomalias.

4.1 Sensor utilizado

O sensor que será utilizado neste projeto, será do tipo LIDAR, modelo M1C1_Mini (Figuras 8 e 9), da fabricante China Science Photon Chip, com capacidade de mapeamento 2D de 100mm a 6000mm em 360º, o que é suficiente para o projeto proposto. Os dados técnicos podem ser vistos na Figura 10 e na Tabela 1.

_

Figura 8: M1C1_mini - Vista superior

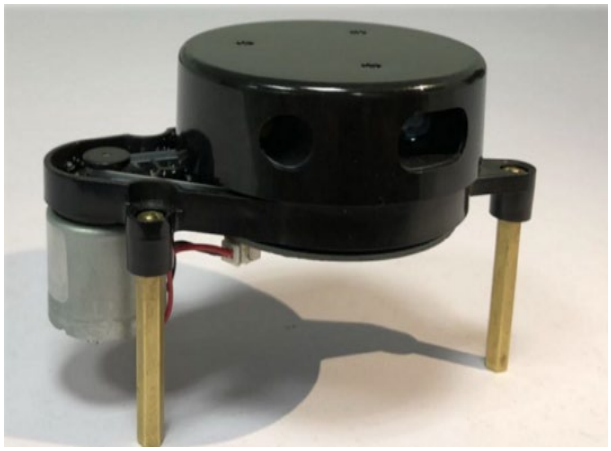

Figura 9: M1C1_mini - Perspectiva

Como o sensor necessita de uma distância mínima de 100mm entre a superfície de disparo do laser e a superfície analisada, era necessária uma tubulação que atendesse esse requerimento técnico, com 300mm de

diâmetro, compensando os 60mm de diâmetro do sensor rotativo e restando 240mm úteis.

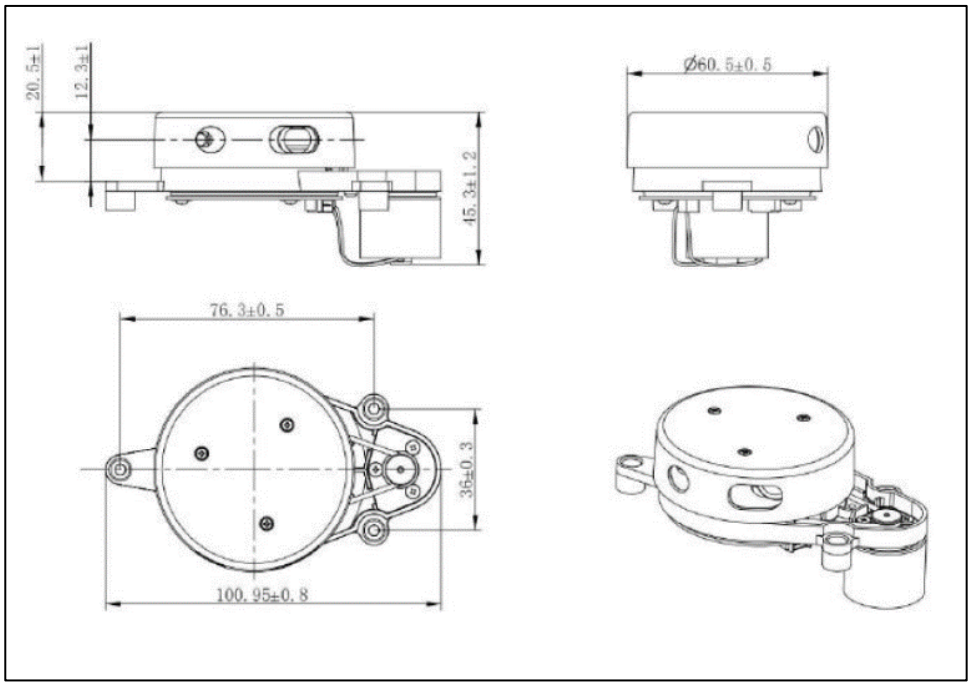

Figura 10: Medidas do Sensor Lidar

Tabela 1: Características do Sensor, retiradas do manual do fabricante.

–	Parameter
Light source	Laser@780nm, Class1
Working principle	Triangulation
Detection distance	0.10~8.0m@90%
Measurement accuracy	mm level@＜1m； 2%@1m-6m
Field of view	360° horizontal
Angle resolution	≈0.93°
Measuring frequency	3860points/s (default)
Rotating speed	10Hz
Power consumption	Typ. 1.0W
Operating Voltage	5V
Dimensions	L100.95 * W60.50 * H45.30mm
Weight	98 g
Communication Interface	UART serial port
Signal format	Angle, distance, etc.

4.2 Aquisição de Dados

A metodologia deste estudo está dividida em duas etapas principais: aquisição e análise de dados, e desenvolvimento do algoritmo de detecção. Na primeira etapa, denominada aquisição e análise de dados, um sensor lidar será posicionado no centro do duto e durante essa fase, foram realizados dois conjuntos de experimentos.

O primeiro conjunto, com experimentos estáticos, o LiDAR foi fixado em uma posição centralizada dentro do duto e foram realizadas leituras da superfície interna do duto sem nenhum furo. Então foram introduzidos poucos furos de diferentes diâmetros na tubulação, com o objetivo de que essas anomalias de geometria e dimensões conhecidas fossem utilizadas como conjunto de treinamento para classificadores supervisionados, permitindo uma detecção gráfica a partir de parâmetros conhecidos dos algoritmos de machine learning e também gerar um conjunto intermediário de dados para treinamento. O sensor realizou varreduras de 360 graus, capturando dados de distância e ângulo em relação aos pontos da superfície interna do duto, onde estas leituras foram registradas e armazenadas em formato digital, permitindo uma análise detalhada posterior.

Figura 12: Conexão USB

Figura 11: Sensor LIDAR - Leitura estática

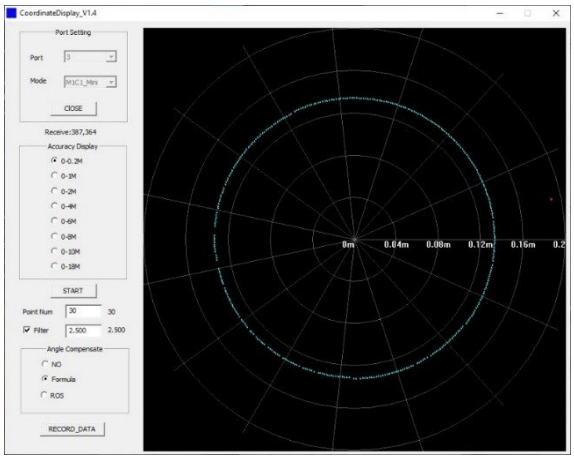

Figura 13: Visualização preliminar do perfil de pontos

Nos experimentos dinâmicos realizados, o sensor LiDAR foi deslocado ao longo do duto, não mantendo-se mais centralizado durante a operação e tendo então ângulo e altura variadas, permitindo avaliar sua influência na qualidade dos dados adquiridos. Durante o movimento, o sensor realizou novamente varreduras de 360 graus, capturando informações de distância e ângulo em relação à superfície interna do duto. As leituras obtidas foram registradas e armazenadas em formato digital novamente, contendo também dados relativos à posição do sensor ao longo do duto. Essa etapa permitiu simular a aplicação do sistema em um ambiente real com variações em relação a posição central, no qual o sensor percorre a tubulação para identificar possíveis derivações clandestinas. E para esta etapa dinâmica, foi utilizado como duto um aro de 300mm de diâmetro produzido em impressora 3D da PUC e contendo furos de diâmetros conhecidos.

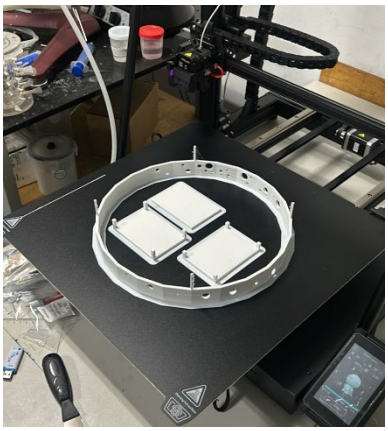

Figura 14: Impressão 3D dos componentes da bancada na PUC-Rio.

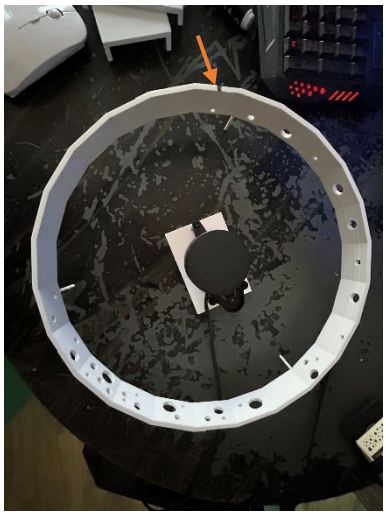

Figura 15: Marcação para variação de ângulo com sensor em suporte para variação da altura em 5mm e posição central sem posição fixa.

Na Figura 15 é possível verificar uma marcação em preto na parte superior do aro, apontado pela seta de cor laranja, esta marcação foi utilizada como posicionamento angular, sendo deslocada de 90º em cada nova tomada de dados pelo sensor LIDAR, de forma que foram utilizadas 3 alturas diferentes, variando 5mm entre cada uma. Estas bases podem ser vistas nas imagens x e y. Desta maneira, o sensor realizou 12 leituras diferentes para esta etapa dinâmica, tendo cada altura quatro medições realizadas, com variação de ângulo e deslocamento da posição central em relação ao aro.

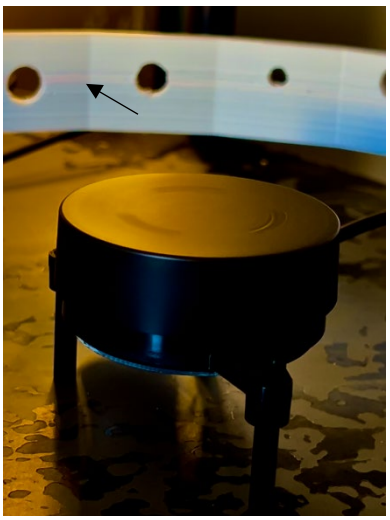

Figura 16:Leitura sendo realizada em altura do próprio sensor

Foram escolhidos furos de 2mm a 8mm pois são medidas inferiores às do padrão de $^{1}/_{2}$" utilizado pelos criminosos nos furtos de combustível em dutos [2]. Essas derivações são, tipicamente, feitas através de perfurações a partir do meio externo. Normalmente, o volume perdido em derivações clandestinas é pequeno e não são facilmente detectáveis por métodos baseados em variação de vazão ou pressão [2].

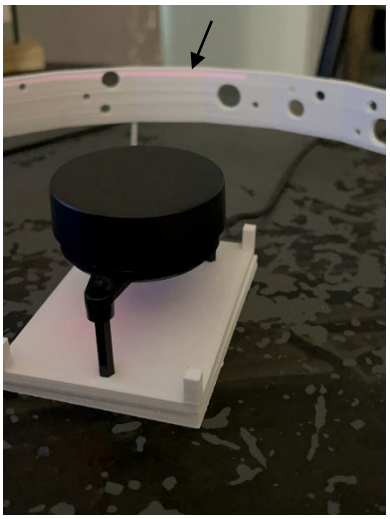

Figura 17: Leitura realizada em última altura com duas bases de 5mm.

Após a conclusão dos experimentos, foram gerados ao todo 14 datasets obtidos das leituras do sensor LIDAR, então os dados coletados, tanto nos testes estáticos quanto nos dinâmicos, foram analisados utilizando técnicas de processamento de sinais e aprendizado de máquina. Essa análise possibilitou a interpretação dos dados e a validação da eficiência do sistema para a detecção de anomalias e derivações na estrutura do duto.

Uma rotina em MATLAB foi escrita para uma análise preliminar e gerados gráficos de uma polar para cada dataset. Além disto, foi necessário realizar um tratamento nos dados de modo a retirar zeros e alguns dados NaN para otimizar o processamento dos algoritmos, reduzindo o tempo necessário para cada algoritmo ser executado. Então foi utilizada a linguagem Python junto com a biblioteca Scikit-Learn de Machine Learning para realizar as análises e gerar os resultados que serão demonstrados a seguir nos próximos tópicos deste trabalho.

5 Resultados

A análise dos resultados obtidos com o sensor LIDAR M1C1_Mini demonstrou a capacidade do sensor de mapear a superfície interna dos dutos, permitindo a identificação de anomalias nas nuvens de pontos, nos casos com tubulações defeituosas.

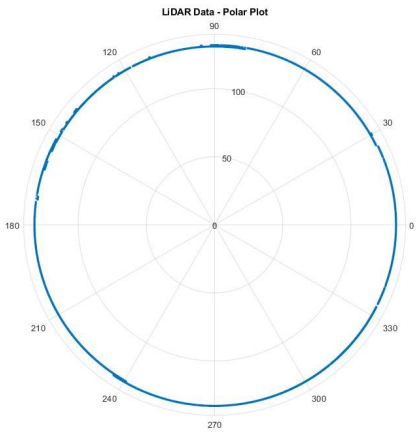

Figura 21: Polar - Sem Furos

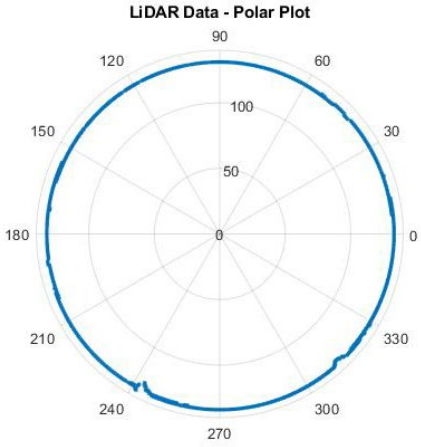

Figura 22: Polar - Poucos Furos

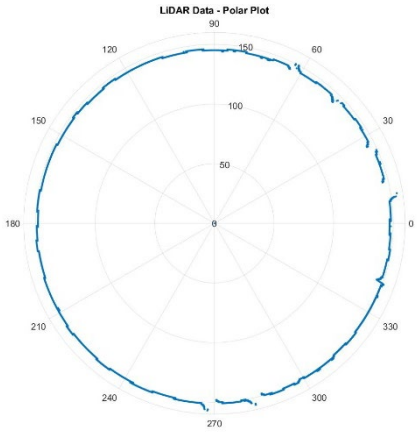

Figura 20:Polar Alt 0 e Pos 1

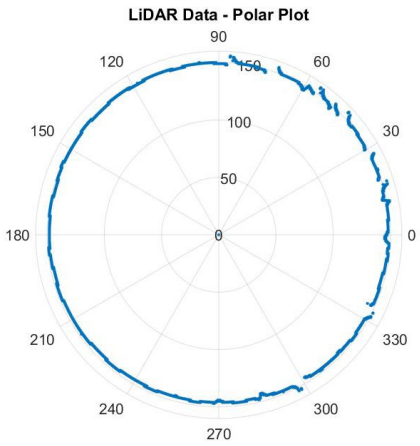

Figura 19: Polar Alt 1 e Pos 1

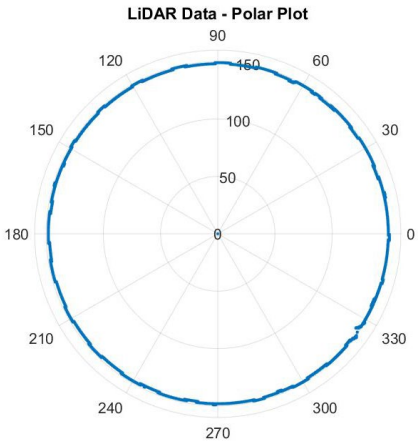

Figura 18: Polar Alt 2 e Pos 1

Nas imagens acima, podemos visualizar algumas das polares obtidas nas leituras do sensor LIDAR. Aqui temos 5 das 14, para verificação visual, sendo as 3 imagens de baixo, uma de cada altura (Alt), onde Alt 0 se refere a altura do próprio sensor, sem estar sobre alguma das bases de 5mm, e todas na posição (Pos) 1 que se refere a posição angular em 0 graus, nas

demais variações o aro é rotacionado de 90º em 90º, demonstrando como o sensor está observando a superfície interna do duto em diferentes alturas e com diferentes furos na trajetória da luz do sensor. Não é necessário colocar todas as polares, tendo em vista que os gráficos de mesma altura têm bastante semelhança, variando claro a posição angular e a posição central levemente deslocada de forma aleatória.

Ao analisar todo o conjunto de dados que estava disponível, vemos na figura a seguir, uma distribuição homogênea das leituras em cada ângulo e também vemos como estão concentradas as distâncias.

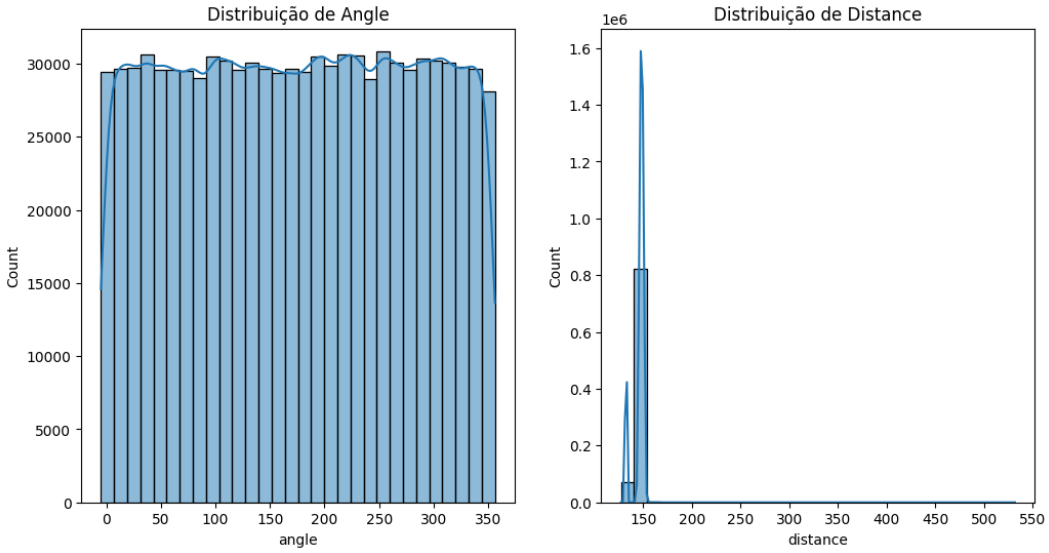

Figura 23: Distribuição de ângulos e distâncias nas leituras.

Foi utilizar no conjunto de treinamento dos modelos, o SMOTE (Synthetic Minority Oversampling Technique**),** que é uma técnica utilizada para lidar com desbalanceamento de classes em problemas de classificação. Ele funciona gerando amostras sintéticas para a classe minoritária, aumentando sua representatividade no conjunto de dados. Em vez de simplesmente replicar as instâncias existentes, o SMOTE cria novas instâncias interpolando os atributos das amostras minoritárias já presentes no conjunto, considerando os k-vizinhos mais próximos. Isso ajuda a evitar problemas de overfitting, comuns em abordagens de duplicação direta de dados minoritários (CHAWLA; BOWYER; HALL; KEGELMEYER, 2002).

Além dos conjuntos de treinamento, posteriormente foram utilizados dois processamentos para gerar a matriz de confusão, o primeiro com os datasets obtidos logo após o tratamento no MATLAB, sem o SMOTE aplicado ao conjunto de treino e outro com a técnica aplicada, que foi especialmente importante neste trabalho para reduzir o tempo de execução dos algoritmos que sem ele estar implementado, chegou a mais de 3 horas de execução em um dos três algoritmos utilizados neste trabalho

A linguagem Python foi utilizada neste trabalho, por se tratar de uma linguagem de programação intuitiva e de fácil implementação, com ampla disponibilidade de conteúdo para aplicação e em especial pela possibilidade de utilizar a biblioteca Scikit-Learn.

Então utilizando a plataforma Google Colab, contando com um ambiente com todas as ferramentas necessárias para implementação de uma solução de Machine Learning com processamento em Cloud e possibilidade de utilização, caso fosse necessário de GPUs do próprio Google,

Foram assim, obtidos os seguintes resultados para cada algoritmo, primeiramente sem a utilização do SMOTE e em seguida com a utilização deste, chegando então aos resultados iniciais e posteriormente também estão expostas as análises e conclusões.

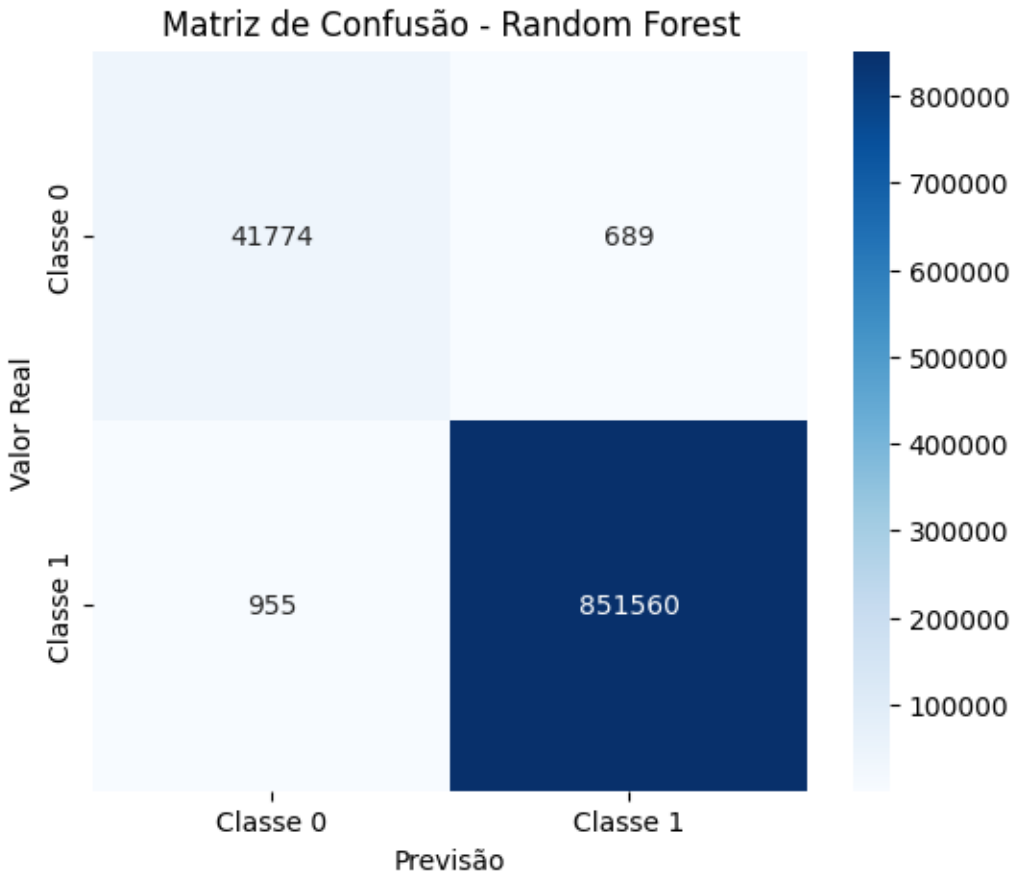

Figura 24: Resultados Random Forest - Sem SMOTE

Observando a matriz, podemos analisar que o modelo classificou corretamente uma grande quantidade de amostras como "com furo", totalizando 851.560 Verdadeiros Positivos (VP). Além disso, classificou corretamente 41.774 amostras como "sem furo", representando os Verdadeiros Negativos (VN). No entanto, o modelo também apresentou 689 Falsos Positivos (FP), ao classificar erroneamente algumas amostras como "com furo" quando, na verdade, eram "sem furo". Por fim, houve 955 Falsos Negativos (FN), em que o modelo classificou equivocadamente algumas amostras como "sem furo" quando, na verdade, eram "com furo".

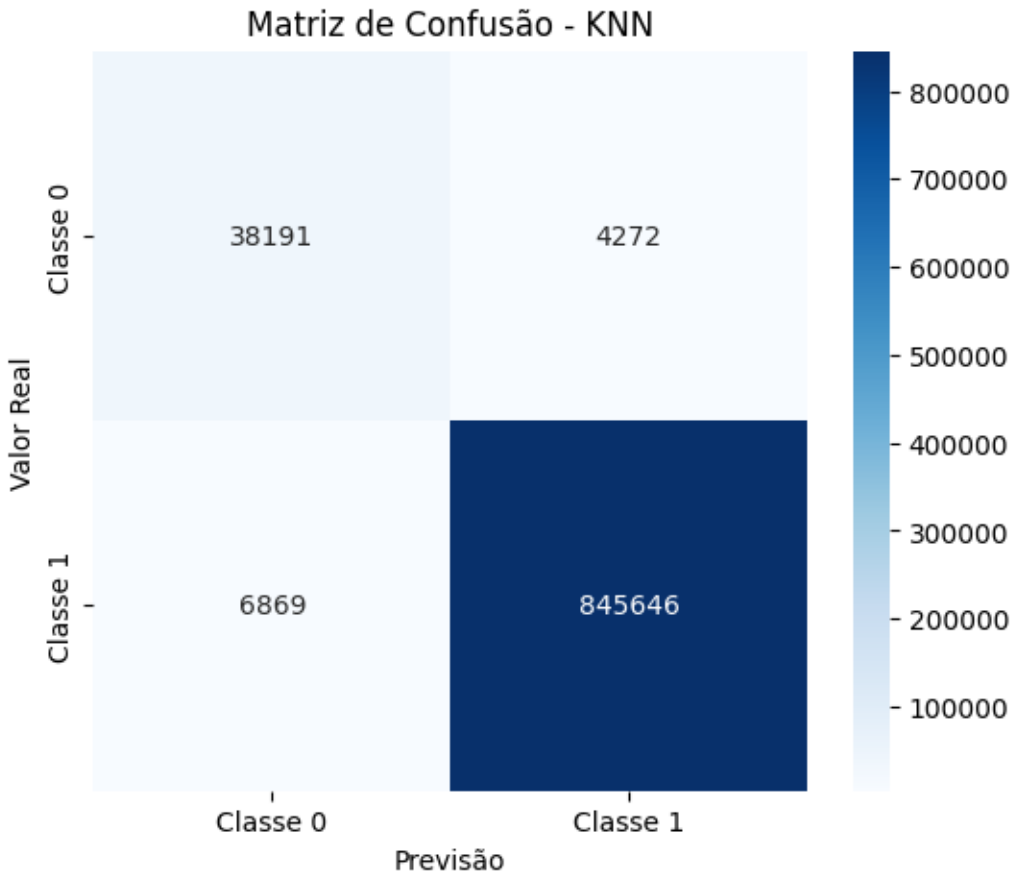

Figura 25: Resultados KNN - Sem SMOTE

Na matriz de confusão do KNN (sem SMOTE), o modelo apresentou 845.646 Verdadeiros Positivos (VP), indicando as amostras corretamente classificadas como "com furo". Foram identificados também 38.191 Verdadeiros Negativos (VN), que representam as amostras corretamente classificadas como "sem furo". Por outro lado, ocorreram 4.272 Falsos Positivos (FP), nos quais o modelo classificou erroneamente amostras como "com furo" quando, na verdade, eram "sem furo". Além disso, registraram-se 6.869 Falsos Negativos (FN), em que o modelo classificou incorretamente amostras como "sem furo" quando, na realidade, pertenciam à classe "com furo".

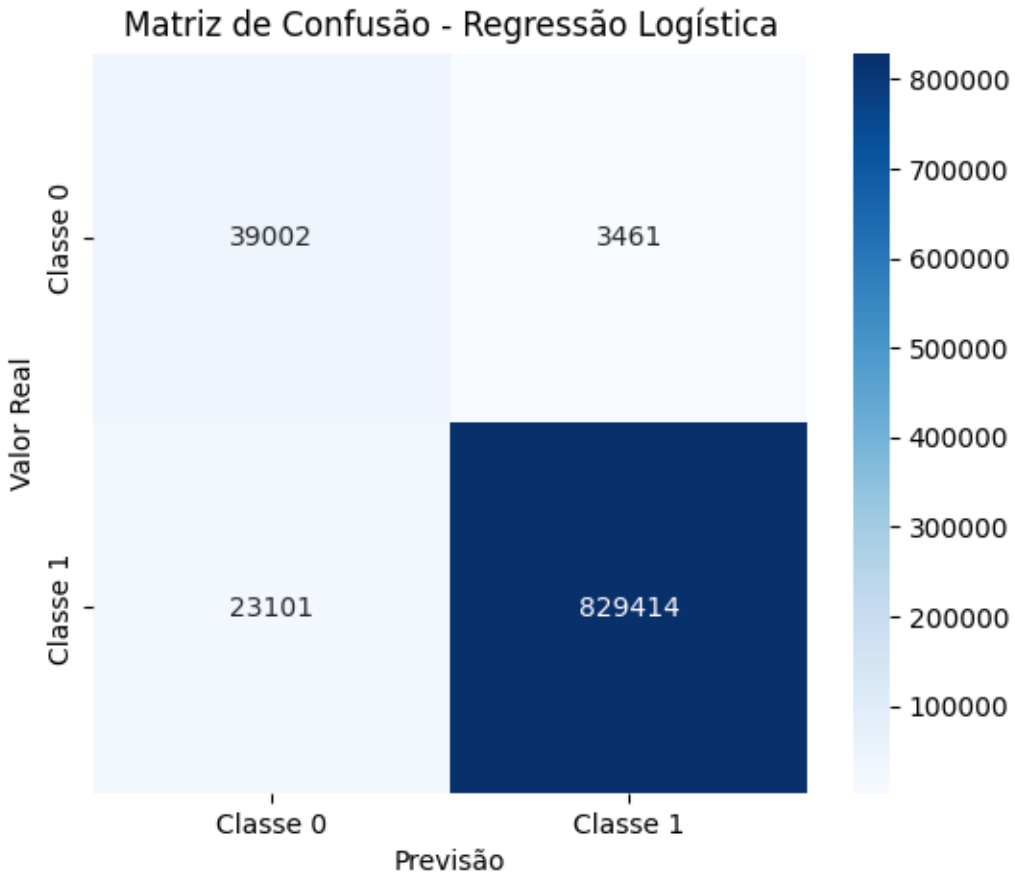

Figura 26: Resultados Logistic Regression - Sem SMOTE

Na matriz de confusão da Regressão Logística (sem SMOTE), o modelo apresentou 829.414 Verdadeiros Positivos (VP), representando as amostras corretamente classificadas como "com furo". Além disso, foram identificados 39.002 Verdadeiros Negativos (VN), que correspondem às amostras corretamente classificadas como "sem furo". Por outro lado, ocorreram 3.461 Falsos Positivos (FP), onde o modelo classificou equivocadamente amostras como "com furo" quando, na realidade, eram "sem furo". Também foram registrados 23.101 Falsos Negativos (FN), em que o modelo erroneamente classificou amostras como "sem furo" quando, na verdade, pertenciam à classe "com furo".

Agora temos os resultados com a técnica SMOTE aplicada ao conjunto de treinamento e perceber o quanto isto influencia na eficiência do modelo.

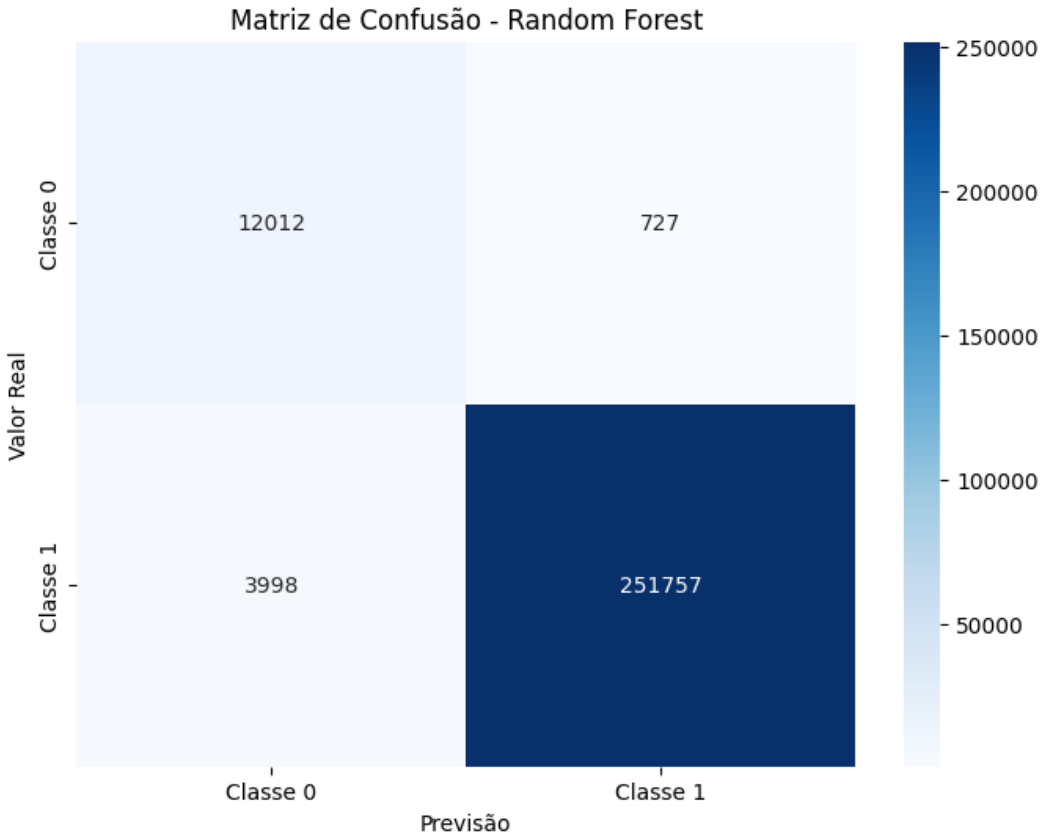

Figura 27: Resultados Random Forest - Com SMOTE

Na matriz de confusão do Random Forest (com SMOTE), o modelo apresentou 251.757 Verdadeiros Positivos (VP), indicando as amostras corretamente classificadas como "com furo". Foram também registrados 12.012 Verdadeiros Negativos (VN), correspondendo às amostras corretamente classificadas como "sem furo". Entretanto, houve 727 Falsos Positivos (FP), nos quais o modelo classificou erroneamente amostras como "com furo" quando, na verdade, eram "sem furo". Além disso, foram identificados 3.998 Falsos Negativos (FN), onde o modelo classificou incorretamente amostras como "sem furo" quando, na realidade, pertenciam à classe "com furo".

Figura 28: Resultados KNN - Com SMOTE

Na matriz de confusão referente ao KNN (com SMOTE), observa-se que o modelo identificou corretamente 251.025 amostras como "com furo", representando os Verdadeiros Positivos (VP). Além disso, foram detectadas 12.690 amostras corretamente classificadas como "sem furo", os Verdadeiros Negativos (VN). No entanto, houve 49 Falsos Positivos (FP), casos em que amostras "sem furo" foram incorretamente atribuídas à classe "com furo". Por outro lado, 4.730 amostras foram tratadas de forma equivocada como "sem furo", constituindo os Falsos Negativos (FN).

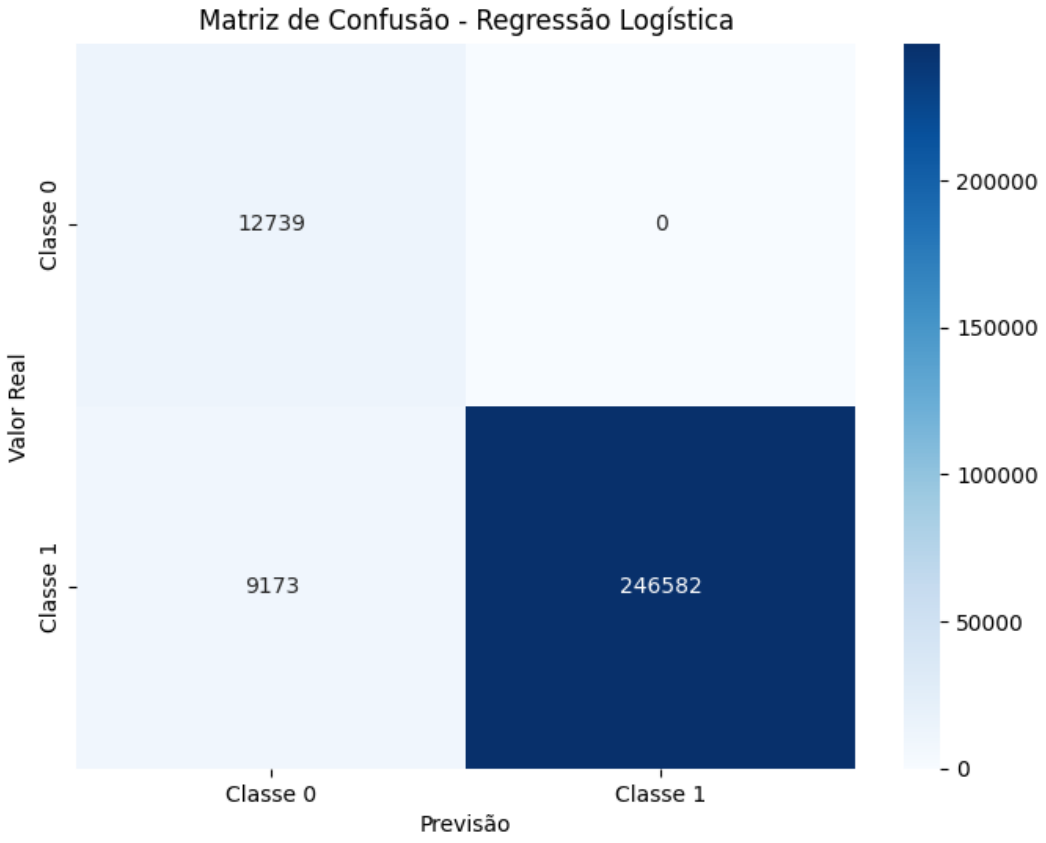

Figura 29:Resultados Logistic Regression - Com SMOTE

Na matriz de confusão da Regressão Logística com SMOTE, nota-se que o modelo conseguiu identificar corretamente 246.582 amostras como "com furo", representando os Verdadeiros Positivos (VP). Além disso, todas as 12.739 amostras "sem furo" foram classificadas corretamente, compondo os Verdadeiros Negativos (VN). Não houve registro de Falsos Positivos (FP), ou seja, nenhuma amostra "sem furo" foi erroneamente atribuída à classe "com furo". Por outro lado, 9.173 amostras foram equivocadamente classificadas como "sem furo", caracterizando os Falsos Negativos (FN).

5.1 Análise do Desempenho Entre Os Algoritmos

Nesta seção é analisado e comparado o desempenho do Random Forest, KNN e Logistic Regression

5.1.1 Matriz De Confusão Completa

A tabela a seguir apresenta a Matriz de Confusão referente ao processamento dos algoritmos Random Forest, KNN e Logistic Regression com e sem SMOTE. Todas as taxas e indicadores de desempenho que serão apresentadas adiante são obtidas a partir desta matriz.

Tabela 2: Matriz de Confusão por Algoritmo

Algoritmo	VP	VN	FP	FN
RF	851560	41774	689	955
KNN	845646	38191	4272	6869
LR	829414	39002	3461	23101
RF - SMOTE	251757	12012	727	3998
KNN - SMOTE	251025	12690	49	4730
LR - SMOTE	246582	12739	0	9173

5.1.2 Taxas de Desempenho

A tabela 3 apresenta os resultados da taxa de acurácia, precisão com e sem furo e recall com e sem furo.

Tabela 3: Acurácia, Precisão e Recall

Algoritmo	Acurácia (%)	Precisão " com furo" (%)	Recall " com furo" (%)	Precisão " sem furo" (%)	Recall " sem furo" (%)
RF	99,82	99,92	99,89	97,76	98,38
KNN	98,76	99,50	99,19	84,76	89,94
LR	97,03	99,58	97,29	62,80	91,85
RF - SMOTE	98,24	99,71	98,44	75,03	94,29
KNN - SMOTE	98,22	99,98	98,15	72,85	99,62
LR - SMOTE	96,58	100,00	96,41	58,14	100,00

A tabela apresenta as métricas de acurácia, precisão e recall para os algoritmos Random Forest (RF), K-Nearest Neighbors (KNN) e Regressão Logística (LR), avaliados com e sem a aplicação do oversampling SMOTE. No caso do RF sem SMOTE, obteve-se a maior acurácia entre todos os modelos (99,82%), com excelente desempenho em precisão e recall para ambas as classes, indicando que o modelo consegue identificar padrões nos dados mesmo sem balanceamento.

Com SMOTE, a acurácia do RF diminuiu para 98,24%, assim como a precisão e o recall para a classe "sem furo", sugerindo que o oversampling pode ter introduzido ruído ou enviesamento, prejudicando a detecção da classe minoritária. O KNN sem SMOTE alcançou a segunda maior acurácia (98,76%), com alta precisão para "com furo" e bom recall para ambas as classes, embora tenha demonstrado menor precisão para "sem furo", refletindo uma dificuldade em identificar essa classe. Com SMOTE, a acurácia permaneceu praticamente inalterada (98,22%), mas houve uma melhora significativa na precisão e recall para "sem furo", mostrando o impacto positivo do balanceamento.

Já a LR sem SMOTE teve a menor acurácia (97,03%), com alta precisão para "com furo", mas baixa para "sem furo", evidenciando um viés para a classe majoritária. Com a aplicação do SMOTE, a acurácia diminuiu para 96,58%, porém houve um aumento expressivo na precisão e recall para "sem furo", com recall atingindo 100%, comprovando a importância do oversampling para equilibrar as classes e melhorar a capacidade do modelo de detectar a classe minoritária. Comparativamente, o RF sem SMOTE apresentou a maior acurácia geral, enquanto o KNN com SMOTE obteve o melhor equilíbrio de desempenho entre as classes. Em termos de recall para "sem furo", a LR com SMOTE destacou-se, alcançando 100%, seguida de perto pelo KNN com SMOTE (99,62%).

5.1.3 Taxas de Desempenho Proporcionais

A tabela 4 apresenta os resultados das taxas de Falso Positivo Relativo (FP_r) e Falso Negativo Relativo (FN_r).

Tabela 4: Média das Taxas de Falso Positivo e Falso Negativo Relativos

Algoritmo	FP_r (%)	FN_r(%)
RF	1,62	0,11
KNN	10,06	0,81
LR	8,15	2,71
RF - SMOTE	5,71	1,56
KNN - SMOTE	0,38	1,85
LR - SMOTE	-	3,59

O Random Forest (RF) destacou-se por apresentar o menor FNr entre todos os algoritmos, tanto com o uso quanto sem o uso do SMOTE, evidenciando excelente desempenho na identificação de casos positivos. Além disso, o FPr do RF é baixo, o que indica que o modelo gera poucos falsos alarmes. O K-Nearest Neighbors (KNN), por sua vez, exibiu um comportamento variado: sem SMOTE, apresentou o maior FPr, indicando que classificou erroneamente muitos casos negativos como positivos. Contudo, com a aplicação do SMOTE, o FPr reduziu-se drasticamente, tornando-se o menor entre todos os algoritmos, embora isso tenha causado um aumento no FNr, comprometendo a identificação de casos positivos.

Já a Regressão Logística (LR) demonstrou um FPr intermediário e um FNr elevado na ausência do SMOTE, o que reflete uma dificuldade significativa em identificar casos positivos. Com o uso do SMOTE, houve uma redução no FNr, mas este ainda permaneceu mais alto em comparação

com o RF, indicando desafios na detecção precisa da classe positiva, ou seja a classe 1, que são os furos.

De forma geral, a aplicação do SMOTE contribuiu para a diminuição do FPr na maioria dos casos, auxiliando na redução de falsos alarmes. No entanto, essa técnica também resultou em um aumento do FNr em alguns algoritmos, com exceção da LR, sugerindo um possível impacto na capacidade dos modelos de identificar casos positivos.

5.1.4 Análise do F-Score

O F-Score é uma métrica amplamente utilizada para avaliar o desempenho de modelos de classificação, especialmente em cenários com classes desbalanceadas. Ele combina a precisão, que representa a proporção de predições corretas entre todas as classificações positivas realizadas pelo modelo, e o recall, que mede a proporção de casos positivos identificados corretamente. Esses dois elementos são combinados por meio de uma média harmônica ponderada, oferecendo uma visão integrada do desempenho do modelo. O parâmetro β ajusta a importância relativa entre precisão e recall. Quando β é igual a 1, como no caso apresentado, a métrica dá igual peso a ambos, tornando-a uma medida balanceada para análise de desempenho (MANNING; RAGHAVAN; SCHÜTZE, 2008) e (WITTEN; FRANK; HALL; PAL, 2016).

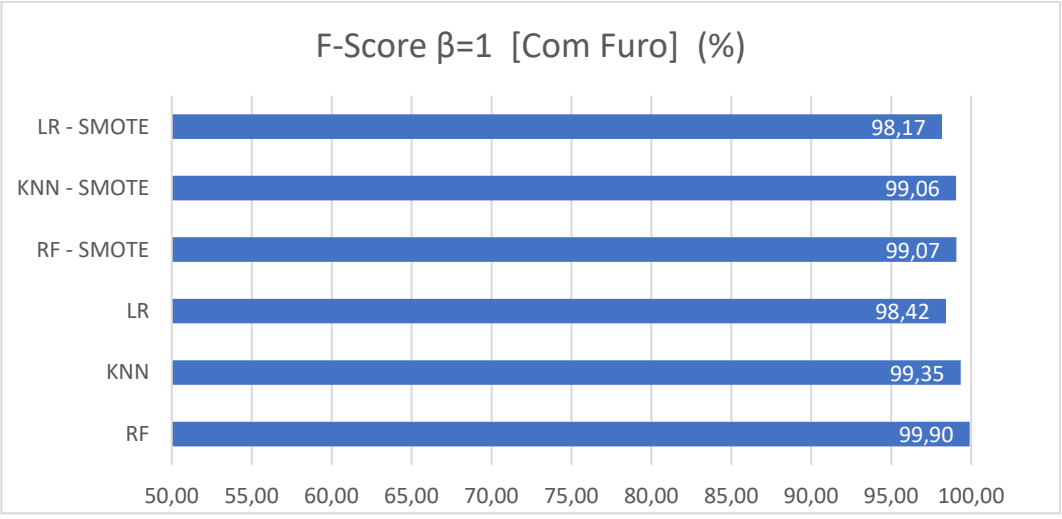

Figura 30: F-Score [%] com β = 1 - Com Furo

No gráfico demonstrado acima, na Figura 30, referente à classe1, "Com Furo", todos os modelos apresentaram desempenho elevado. O RF destacou-se novamente como o melhor modelo, obtendo 99,90% sem SMOTE e 99,07% com SMOTE, evidenciando excelente precisão e recall para esta classe. O KNN também obteve valores altos, com 99,35% sem SMOTE e 99,06% com SMOTE. Já a LR alcançou 98,42% sem SMOTE e 98,17% com SMOTE, apresentando um desempenho levemente inferior em comparação aos demais.

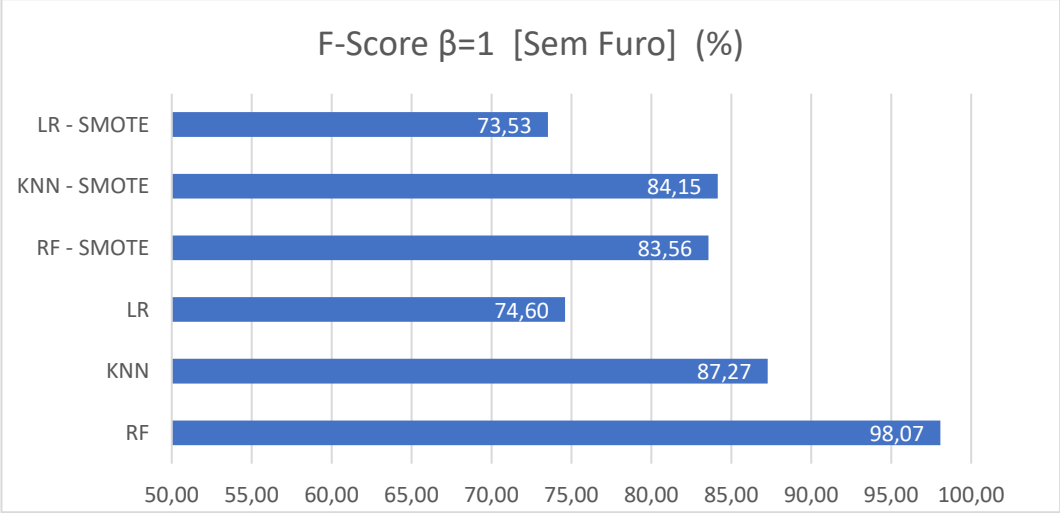

Figura 31: F-Score [%] com β = 1 - Sem Furo

Ao observar a Figura 31, vemos que o F-Score para a classe 0, "Sem Furo", o modelo Random Forest (RF) sem SMOTE obteve o melhor desempenho (98,07%), seguido pelo K-Nearest Neighbors (KNN) sem SMOTE (87,27%). Isso demonstra que esses modelos têm maior capacidade de classificar corretamente as amostras desta classe. A aplicação do SMOTE para o RF resultou em uma leve diminuição no F-Score (83,56%), indicando que o oversampling introduziu algum ruído. O KNN, com SMOTE, melhorou significativamente para 84,15%, evidenciando o impacto positivo da técnica no balanceamento das classes. Já a Regressão Logística (LR) obteve os menores valores, tanto com (73,53%) quanto sem SMOTE (74,60%), refletindo dificuldades em identificar corretamente os casos "Sem Furo".

6 Conclusão

No presente trabalho, foi realizado um estudo da detecção de anomalias em dutos utilizando um sensor LiDAR 2D, onde os resultados mostraram que o sensor foi capaz de fornecer dados para a identificação de pequenos furos em uma tubulação de 300mm de diâmetro. Utilizou-se algoritmos de machine learning para a identificação de conjuntos de dados com e sem a presença de furos nos dutos. Os algoritmos foram capazes de detectar anomalias com taxas de acerto acima de 90%.

Na análise comparativa dos algoritmos, o Random Forest (RF) sem SMOTE demonstrou ser o modelo mais eficiente em termos de desempenho geral. Apresentou alta acurácia mesmo em cenários com classes desbalanceadas. Por outro lado, o K-Nearest Neighbors (KNN) com SMOTE destacou-se por alcançar a maior precisão para a classe minoritária ("sem furo"), sendo particularmente útil em situações em que a detecção dessa classe é prioritária.

O uso da técnica de oversampling SMOTE revelou um impacto positivo na performance dos modelos KNN e Regressão Logística (LR), ao melhorar a identificação da classe "sem furo". Contudo, esse benefício veio acompanhado de um custo em termos de acurácia geral, o que reforça a necessidade de considerar as prioridades específicas do problema ao selecionar o modelo ideal, assim, a escolha do modelo final depende das demandas do contexto, se a minimização de falsos negativos é crucial ou não, como neste caso é, assim o RF se apresenta como a melhor escolha devido ao seu baixo FNr.

Logo, em resumo, o RF destacou-se como o modelo com melhor desempenho geral, mantendo o menor FNr em ambos os cenários (com e sem SMOTE). O KNN com SMOTE mostrou-se eficaz em minimizar o FPr, embora com um aumento no FNr. Já a LR apresentou resultados intermediários, enfrentando dificuldades consideráveis, especialmente na identificação de casos positivos, mesmo com o auxílio do SMOTE. Assim, a escolha do modelo deve ser orientada pelas prioridades específicas do

problema em questão, equilibrando a acurácia geral e a performance nas classes de maior interesse.

O modelo de detecção abordado no trabalho, juntamente com o processamento de dados escolhido, alcança uma precisão de detecção elevada quando comparado a outros métodos como inspeção visual e análise de vibrações, que não detectariam os orifícios testados neste projeto, sendo superior inclusive a outros estudos envolvendo análise de imagens oriundas de câmeras (OBAID M H; HAMAD, A H, 2023)

Apesar dos desafios, o uso deste tipo de sensor, é uma tecnologia promissora na detecção de anomalias em dutos. Com o uso de sensores mais precisos e robustos além de métodos de processamento de imagens mais refinados, o uso de sensores LiDAR pode se tornar a solução interessante para a detecção de anomalias em tubulações. A principal contribuição deste trabalho foi avaliar uma prova de conceito de uso da tecnologia em cenários com furos menores do que aqueles esperados em aplicações práticas. Outro ponto importante abordado aqui foi a avaliação da acurácia na detecção de anomalias mesmo com o uso de um sensor de baixo custo e de resolução espacial da mesma ordem de grandeza das anomalias.

No entanto, ainda há espaço para melhorias no método aqui desenvolvido e executado. Uma possível melhoria é a utilização destas e outras técnicas de aprendizado não supervisionado aliados a utilização de sensores LIDAR com maior resolução espacial, incluindo sensores 3D. Além disso, o treinamento da rede para a identificação e classificação das anomalias pode ser feito em um estudo futuro.

7 Referências

[1] Agência Nacional do Petróleo, Gás Natural e Biocombustíveis (ANP). Anuário Estatístico 2022. Brasília, DF, ANP, 2023.

SOARES, Gabriel Antonio Silva; GALVÍNCIO, Josiclêda Domiciano. Uso do LiDAR para avaliar os padrões hídricos de bacias em áreas urbanas: Caracterização fisiográfica da bacia do Rio Beberibe, PE. Revista Brasileira de Geografia Física, v. 13, n. 7, p. 3659-3674, 2020. DOI: http://dx.doi.org/10.26848/rbgf.v13.07.p3659-3674.

LIU, C.-W. et al. A new leak location method based on leakage acoustic waves for oil and gas pipelines. Journal of Loss Prevention in the Process Industries, OXFORD, v. 35, p. 236-246, 2015. ISSN 0950-4230. Doi: http://dx.doi.org/10.1016/j.jlp.2015.05.006

LIU, H. Pipeline engineering. CRC Press, 2003. ISBN 0203506685.

Valery M. Petoukhov, Zaytuna K. Petoukhova, Rishad A. Akhtiamov, German Ivanovych Il'in, Oleg G. Morozov, and Yuri E. Pol'ski "Lidar technologies application to leakage detection in oil product pipelines", Proc. SPIE 3588, Nondestructive Evaluation of Utilities and Pipelines III, (5 February 1999); https://doi.org/10.1117/12.339944

GONG, Jie. Mobile Hybrid LiDAR & Infrared Sensing for Natural Gas Pipeline Monitoring. New Brunswick, NJ: Rutgers University. Center for Advanced Infrastructure and Transportation, 2016. (RITARS-14-H-RUT). Disponível em: https://rosap.ntl.bts.gov/view/dot/32121.

[2] JEONG, Soonho; KIM, Jinseok; YOU, Soohyun. Wireless Portable LDS for Theft Detection. In: PIPELINE TECHNOLOGY CONFERENCE, 2017, at Berlin, Germany, 2017. Disponível em: https://www.researchgate.net/publication/317303941_Wireless_Portable_LDS_for_Theft_Detection

CAMERINI, Daniel Almeida. Desenvolvimento de Pigs Instrumentados para Detecção e Localização de Pequenos Vazamentos em Dutos. Rio de Janeiro, Setembro, 2004. Disponível em: https://www.maxwell.vrac.puc-rio.br/est_conteudo.php?nrSeq=6229@1

COLOMBAROLI, Pedro Lucio Stefani; BORTONI, Edson da Costa; MARTINS, Helga Gonzaga. Sistema de Detecção de Vazamento em Dutos de Petróleo. In: CONGRESSO BRASILEIRO DE PESQUISA E DESENVOLVIMENTO EM PETRÓLEO E GÁS, 5., 2009, Fortaleza: ABPG, 2009. Disponível em: https://www.yumpu.com/pt/document/view/12793248/sistema-de-deteccao-de-vazamento-em-dutos-de-petroleo-anp

GEREMIA, Giovani. Sistema Autônomo de Inspeção de Dutos. 2012. Dissertação (Mestrado em Engenharia) - Programa de Pós-Graduação em Engenharia de Minas, Metalúrgica e de Materiais, Universidade Federal do Rio Grande do Sul, Porto Alegre, 2012. Disponível em: https://lume.ufrgs.br/handle/10183/72898

OLIVEIRA, Laíse Ramonny Nunes de; FARIAS, Larissa Freitas; CABRAL, Endyara de Morais. Principais técnicas de monitoramento de vazamentos causados por corrosão em dutos de petróleo - Revisão. In: CONGRESSO NACIONAL DE ENGENHARIA DE PETRÓLEO, GÁS NATURAL E BIOCOMBUSTÍVEIS, 2. WORKSHOP DE ENGENHARIA DE PETRÓLEO, Campina Grande: CONEPETRO, 2016. Disponível em: https://editorarealize.com.br/index.php/artigo/visualizar/27026

[3] BASTOS, D. G. D.; NASCIMENTO, P. S.; LAURETTO, M. S. Proposta e análise de desempenho de dois métodos de seleção de características para random forests. SBSI Sociedade Brasileira de Sistemas, Maio 2013. DOI: https://doi.org/10.5753/sbsi.2013.5675

GIONGO, M.; KOEHLER, H. S.; MACHADO, S. do A.; KIRCHNER, F. F.; MARCHETTI, M. LiDAR: princípios e aplicações florestais. Pesquisa Florestal Brasileira, [S. l.], v. 30, n. 63, p. 231, 2010. DOI: 10.4336/2010.pfb.30.63.231. Disponível em: https://pfb.cnpf.embrapa.br/pfb/index.php/pfb/article/view/148. Acesso em: 18 jun. 2023.

SANTOS, Judá Teixeira et al. Desenvolvimento de instrumentação para geração de nuvem de pontos usando sensores inerciais e LiDAR. In: SIMPÓSIO BRASILEIRO DE AUTOMAÇÃO INTELIGENTE, 14., 27 a 30 Out. 2019, Ouro Preto-MG. Anais [...], Campinas, Galoá, 2019. v. 1, p. 1901-1907, 2019-108332. Disponível em: http://www.repositorio.ufc.br/handle/riufc/64719. Acesso em: 18 jun. 2023.

EFRON, B. Bootstrap methods: Another look at the jackknife. The Annals of Statistics, v. 7, p. 1–26, 1979.Disponível em: http://www.jstor.org/stable/2958830

EFRON, B.; TIBSHIRANI, R. J. An Introduction to the Bootstrap. [S.l.]: CHAPMAN HALL, 1993.

[4] BREIMAN, L. Bagging predictors. Machine Learning, v. 24, p. 123–140, 1996.

[5] BREIMAN, L. Random Forests. Machine Learning, v. 45, n. 1, p. 5-32, 2001.

CRIMINISI, A.; SHOTTON, J.; KIPMAN, A. Decision Forests for Classification, Regression, Density Estimation, Manifold Learning and Semi-Supervised Learning. Microsoft Research Cambridge, Technical Report, 2012. Disponível em: https://www.microsoft.com/en-us/research/wp-content/uploads/2016/02/decisionForests_MSR_TR_2011_114.pdf

RODRIGUEZ-GALIANO, V. et al. An assessment of the effectiveness of a random forest classifier for land-cover classification. ISPRS Journal of Photogrammetry and Remote Sensing, v. 67, p. 93-104, 2012.

STROBL, C. et al. Bias in random forest variable importance measures: Illustrations, sources and a solution. BMC Bioinformatics, v. 8, n. 1, p. 25, 2007.

CHAWLA, N. V.; BOWYER, K. W.; HALL, L. O.; KEGELMEYER, W. P. SMOTE: Synthetic Minority Over-sampling Technique. Journal of Artificial Intelligence Research, v. 16, p. 321-357, 2002. DOI: https://doi.org/10.1613/jair.953.

LU, Hongfang; ISELEY, Tom; BEHBAHANI, Saleh; FU, Lingdi. Leakage detection techniques for oil and gas pipelines: State-of-the-art. Tunnelling and Underground Space Technology, v. 98, p. 103249, 2020. Disponível em: https://doi.org/10.1016/j.tust.2019.103249.

BASTOS, D.G.; NASCIMENTO, P.S.; LAURETTO, M.S. Análise empírica de desempenho de quatro métodos de seleção de características para random forests. Revista Brasileira de Sistemas de Informação 7(2), 2014. Disponível em: http://www.seer.unirio.br/index.php/isys/article/view/3309

Frizzarini, C. Algoritmo para indução de árvores de classificação para dados desbalanceados. Dissertação de Mestrado. Programa de Pós-Graduação em Sistemas de Informação, Universidade de São Paulo, 2013. Disponível em: https://teses.usp.br/teses/disponiveis/100/100131/tde-19022014-101043/publico/ClaudioFrizzarini.pdf

GENUER, R.; Poggi, J-M. Random Forests with R. Use R! Series, Cham: Springer, 2020.

HASTIE, T.; TIBSHIRANI, R.; FRIEDMAN (2009), J. The Elements of Statistical Learning. 2ndEdition. Springer.

HESTERBERG, T. et al. Bootstrap Methods and Permutation Tests. Companion Chapter 18 to The Practice of Business Statistics. 2003 https://www.chrisbilder.com/boot/schedule/boot_intro_pbs18.pdf

MANNING, C. D.; RAGHAVAN, P.; SCHÜTZE, H. *Introduction to Information Retrieval*. Cambridge: Cambridge University Press, 2008. Disponível em: https://nlp.stanford.edu/IR-book/. Acesso em: 25 nov. 2024.

WITTEN, I. H.; FRANK, E.; HALL, M. A.; PAL, C. J. *Data Mining: Practical Machine Learning Tools and Techniques*. 4. ed. Cambridge: Morgan Kaufmann, 2016. Disponível em: https://www.sciencedirect.com/book/9780128042915/data-mining. Acesso em: 25 nov. 2024.

[6] COVER, T.; HART, P. Nearest neighbor pattern classification. IEEE Transactions on Information Theory, v. 13, n. 1, p. 21-27, 1967. Doi: https://doi.org/10.1109/TIT.1967.1053964

[7] CUNNINGHAM, P.; DELANY, S. J. k-Nearest neighbour classifiers. Multiple Classifier Systems, p. 1-17, 2007. Doi: http://dx.doi.org/10.1145/3459665

[8] WU, X.; KUMAR, V.; QUINLAN, J. R.; GHOSH, J.; YANG, Q.; MOTODA, H.; MCLACHLAN, G. J.; NG, A.; LIU, B.; YU, P. S.; ZHOU, Z.-H. Top 10 algorithms in data mining. Knowledge and Information Systems, v. 14, n. 1, p. 1-37, 2008.Doi:http://dx.doi.org/10.1007/s10115-007-0114-2

[9] ALPAYDIN, E. Introduction to machine learning. MIT press, 2020.

[10] JAMES, G.; WITTEN, D.; HASTIE, T.; TIBSHIRANI, R. An introduction to statistical learning. Springer, 2013.

[11] BIAU, G.; DEVROYE, L. Lectures on the nearest neighbor method, 2015. Doi: http://dx.doi.org/10.1007/978-3-319-25388-6

IMANDOUST, Sadegh Bafandeh et al. Application of k-nearest neighbor (knn) approach for predicting economic events: Theoretical background. International journal of engineering research and applications, v. 3, n. 5, p. 605-610, 2013. Disponível em: https://www.researchgate.net/publication/304826093_Application_of_K-nearest_neighbor_KNN_approach_for_predicting_economic_events_theoretical_background

[12] S. Zhang, X. Li, M. Zong, X. Zhu and R. Wang, "Efficient kNN Classification With Different Numbers of Nearest Neighbors," in *IEEE Transactions on Neural Networks and Learning Systems*, vol. 29, no. 5, pp. 1774-1785, May 2018, doi: https://doi.org/10.1109/TNNLS.2017.2673241.

[13] CHOMBOON, Kittipong et al. An empirical study of distance metrics for k-nearest neighbor algorithm. In: Proceedings of the 3rd international conference on industrial application engineering. 2015. p. 4. Doi: http://dx.doi.org/10.12792/iciae2015.051

SOUZA, Rodrigo Buchfink de. Desenvolvimento de elementos cerâmicos para uso em dispositivos de inspeção de dutos (PIGS). 2010. Disponível em: https://lume.ufrgs.br/handle/10183/26490

CARBALLO, M.T. (2002) Predição da Macrossomia Fetal Através da Regressão Logística e de Redes Neurais Artificiais. Monografia (Bacharelado em Estatística), Universidade Federal do Rio Grande do Sul - UFRGS, Porto Alegre. Disponível em: https://lume.ufrgs.br/handle/10183/130795

[14] HOSMER JR, D. W.; LEMESHOW, S.; STURDIVANT, R. X. Applied logistic regression. John Wiley & Sons, 2013.

[15] KLEINBAUM, D. G.; KLEIN, M. Logistic regression: A self-learning text. Springer, 2002.

[16] PENDERGAST, J. F.; GASKIN, D. J.; NEWTON, M. A. A survey of methods for analyzing clustered binary response data. International Statistical Review, v. 64, n. 1, p. 89-118, 1996. Doi: https://doi.org/10.2307/1403425

[17] COX, D. R. (1958). The Regression Analysis of Binary Sequences. Journal of the Royal Statistical Society. Series B (Methodological), 20(2), 215–242. Disponível em: http://www.jstor.org/stable/2983890

OBAID, Muhammad H.; HAMAD, Ali H. Deep Learning Approach for Oil Pipeline Leakage Detection Using Image-Based Edge Detection Techniques. Journal Européen des Systèmes Automatisés, v. 56, n. 4, p. 1-10, 2023. DOI: https://doi.org/10.18280/jesa.560416.

FERREIRA, G. T. SISTEMA DE MAPEAMENTO TRIDIMENSIONAL DE AMBIENTES – Trabalho de Conclusão de Curso – Escola de Engenharia de São Carlos - Universidade de São Paulo, São Paulo, 2014.

BIDA, Alexandre; PADOVEZI JUNIOR, Carlos Alberto Borges; CHRISTOFF, Paulo Rycardo Teodoro. Mapeamento de ambientes com LIDAR. 2020. Trabalho de conclusão de curso (Bacharelado em Engenharia da Computação) – Centro Universitário Internacional Uninter, Curitiba, 2020.

NGUYEN H-H, PARK J-H, JEONG H-Y. A Simultaneous Pipe-Attribute and PIG-Pose Estimation (SPPE) Using 3-D Point Cloud in Compressible Gas Pipelines. *Sensors*. 2023; 23(3):1196. doi: https://doi.org/10.3390/s23031196

MAIA, D. M.; MENDES, J. V. S.; SILVA, J. P. A. M.; BASTOS, R. F.; SILVA, M. dos S.; MIRRE, R. C.; MELO, T. R. de; LEPIKSON, H. A. IoT Leak Detection System for Onshore Oil Pipeline Based on Thermography. Sensors, v. 24, n. 21, p. 6960, 2024. Doi: https://doi.org/10.3390/s24216960