UNIVERSIDADE FEDERAL DE ITAJUBÁ

PROGRAMA DE PÓS-GRADUAÇÃO EM ENGENHARIA DE PRODUÇÃO

Modelo preditivo do teor de óleos e graxas em água produzida quantificado pelo método gravimétrico

Simone Carneiro Streitenberger

Itajubá

Junho de 2022

Simone Carneiro Streitenberger

Modelo preditivo do teor de óleos e graxas em água produzida quantificado pelo método gravimétrico

Tese submetida ao Programa de Pós-Graduação em Engenharia de Produção da Universidade Federal de Itajubá, como parte dos requisitos para a obtenção do título de Doutor em Ciências em Engenharia de Produção.

Área de concentração: Engenharia de Produção

Orientador: Prof. Anderson Paulo de Paiva, Dr.

Coorientador: Aloisio Euclides Orlando Jr., Dr.

Itajubá

Junho de 2022

UNIVERSIDADE FEDERAL DE ITAJUBÁ

PROGRAMA DE PÓS-GRADUAÇÃO EM ENGENHARIA DE PRODUÇÃO

Simone Carneiro Streitenberger

Modelo preditivo do teor de óleos e graxas em água produzida quantificado pelo método gravimétrico

Tese aprovada por banca examinadora no dia 10 de junho de 2022, conferindo à autora o título de Doutora em Ciências em Engenharia de Produção.

Banca examinadora:

Prof. Dr. Claudimar Pereira da Veiga (UFPR)

Prof. Dr. Wesley Vieira da Silva (UFAL)

Prof. Dr. Pedro Paulo Balestrassi (UNIFEI)

Prof. Dr. Antonio Fernando Branco Costa (UNIFEI)

Dr. Aloisio Euclides Orlando Jr. (Coorientador - PETROBRAS)

Prof. Dr. Anderson Paulo de Paiva (Orientador)

Itajubá

Junho de 2022

DEDICATÓRIA

Aos meus pais, por estarem por mim sempre e incondicionalmente.

AGRADECIMENTOS

Agradeço a Deus, e à minha família pelo apoio incondicional e por serem minha segurança.

Ao meu orientador, Prof. Dr. Anderson Paiva, pela amizade que construímos e pela compreensão e generosidade com que me conduziu, me ensinando mais do que teorias acadêmicas.

Aos professores do Instituto de Engenharia de Produção e Gestão por terem contribuído tanto para meu crescimento, em especial ao professor Dr. Pedro Paulo Balestrassi pela oportunidade que me ofereceu, pelo cuidado e por todo ensinamento.

Ao meu companheiro, Frank, por ter compartilhado comigo os dias dessa caminhada árdua e importante.

Aos tantos colegas e amigos conquistados durante esta trajetória, por todo carinho e suporte que me deram, imprescindíveis para que eu chegasse até aqui. Em especial aos amigos Aline, Kelly, Andreza, Fabrício, Alexandre e Vinicius. Ao meu grande e querido amigo e companheiro incansável de estudos, Estevão, por ter estado comigo de perto ou de longe: me ensinando, contribuindo para minha evolução, enfrentando dificuldades, comemorando conquistas e, principalmente, não me deixando esquecer a essência da vida.

A UNIFEI, aos órgãos de fomento à pesquisa FAPEMIG, CNPq e especialmente à CAPES e à Petrobras pelo apoio financeiro e pela oportunidade de desenvolver um trabalho com tamanha relevância.

RESUMO

A água produzida gerada pelo processamento primário de petróleo realizado por plataformas petrolíferas localizadas no oceano (*offshore*), que possui um total de óleo e graxas (TOG), geralmente é reinjetada ou descartada no mar. Este descarte é monitorado por órgãos reguladores ambientais que determinam valores máximos de TOG. No Brasil, o método homologado para a medição de TOG é o gravimétrico, que deve ser realizado em laboratórios em terra. Devido à logística de transferência das amostras da plataforma para o laboratório, o resultado da medição é disponibilizado aproximadamente 20 dias após o dia da coleta. Este trabalho propõe o desenvolvimento de um modelo preditivo de TOG gravimétrico (TOG-G) a partir de variáveis de processo, adicionadas de uma variável extraída da variável de resposta, que possa ser utilizado *offshore* e em tempo real, para orientar de maneira mais ágil possíveis ações preventivas ou corretivas a fim de evitar seu desenquadramento. Para isto, as observações foram agrupadas em classes associadas a faixas de TOG-G, por meio das quais foi realizado o balanceamento da base. Conjuntos de treinamento e teste foram gerados e construiu-se um classificador para o agrupamento em função das variáveis de processo mais significativas para a previsão de TOG-G, identificadas por meio de regressão linear. Na sequência, o TOG-G foi modelado a partir das variáveis de processo significativas e do agrupamento. Os resultados obtidos para o conjunto de teste foram avaliados por meio das métricas Erro Médio Absoluto (MAE), Erro Percentual Absoluto Médio (MAPE), coeficiente de determinação (R^2) e coeficiente de correlação de Pearson (ρ), e se mostraram superiores tanto às previsões geradas pelo modelo de previsão desenvolvido a partir dos mesmos previsores, mas desconsiderando o agrupamento, quanto aos valores reais das medições de TOG espectrofotométrico (TOG-S), que constitui o método de tempo real atualmente utilizado como referência na plataforma. Para validação dos ganhos de acurácia com o método proposto, este foi também aplicado a um conjunto clássico de regressão linear de previsão do peso de peixes. Assim, a inclusão da informação do agrupamento ao modelo do TOG-G mostrou-se uma abordagem inovadora e eficiente para aumentar a acurácia de sua previsão a partir de informações disponíveis na plataforma, o que pode beneficiar consideravelmente a indústria petrolífera em termos de controle do processo.

Palavras-chave: Teor de óleos e graxas; Método de previsão; Métodos de classificação; Regressão linear múltipla.

ABSTRACT

The produced water generated by the primary oil processing carried out by offshore oil platforms, which has a total of oil and grease (TOG), is usually reinjected or disposed into the open ocean. This disposal is monitored by environmental regulatory agencies that determine maximum TOG values. In Brazil, the gravimetric method is that homologated for measuring TOG, which must be carried out in onshore laboratories. Due to the logistics of transferring samples from the platform to the laboratory, the measurement result is available approximately 20 days after the day of collection. This work proposes the development of a predictive model of gravimetric TOG (TOG-G) from process variables, added to a variable extracted from the response variable, which can be used offshore and in real time, to more quickly guide possible preventive or corrective actions in order to avoid its non-compliance. For this, the observations were grouped into clusters associated with TOG-G ranges, through which the base balancing was performed. Training and test sets were generated and a classifier was built for the cluster according to the most significant process variables for the prediction of TOG-G, identified through linear regression. Subsequently, the TOG-G was modeled from the significant process variables and the cluster. The results obtained for the test set were evaluated by means of Mean Absolute Error (MAE), Mean Absolute Percent Error (MAPE), coefficient of determination (R^2) and Pearson's correlation coefficient (ρ), and showed to be superior both to the forecasts generated from the predictive model developed from the same forecasters, but disregarding the cluster, as to the real values of the spectrophotometric TOG (TOG-S) measurements, which constitutes the real-time method currently used as a reference in the platform. To validate the gains in accuracy with the proposed method, it was also applied to a classical set of linear regression for predicting fish weight. Thus, the inclusion of the cluster information in the TOG-G model proved to be an innovative and efficient approach to increase the accuracy of its prediction from information available on the platform, which may considerably benefit the oil industry in terms of process control.

Keywords: *Total oil and grease; Prediction method; Classification methods; Multiple Linear Regression*

.

LISTA DE FIGURAS

Figura 1 - Plataforma FPSO (Fonte: (Agência Brasil, 2017)) 23
Figura 2 - Exemplo de posicionamento de hidrociclones e flotador em uma parte do processo de tratamento de água de uma FPSO (Fonte: autoria própria) 24
Figura 3 – Estrutura de uma árvore de decisão de 2 níveis (Fonte: autoria própria)..... 29
Figura 4 - Ilustração do conceito de bagging (Fonte: adaptado de Sen, 2021) 29
Figura 5 - Exemplo de classificação via KNN (Fonte: adaptado de Müller e Guido (2016)) 31
Figura 6 - Exemplo de uma ANN com uma camada escondida (Fonte: adaptado de Balestrassi et al., 2009)..... 32
Figura 7 - Tipos de regressão logística (Fonte: autoria própria) 33
Figura 8 - Exemplo de classificação via SVM (Fonte: adaptado de Müller e Guido (2016)) 35
Figura 9 - Composição da base fiscal (Fonte: autoria própria) 37
Figura 10 - Composição da base diária a partir da base fiscal (Fonte: autoria própria). 38
Figura 11 - Fluxograma da Etapa de Classificação – A (Fonte: autoria própria)..... 41
Figura 12 - Fluxograma da Etapa de Previsão – B (Fonte: autoria própria) 44
Figura 13 - Boxplot do TOG-G (Fonte: autoria própria) 52
Figura 14 - Conjunto de dados desbalanceado (a) e balanceado (b) através do undersampling (Fonte: autoria própria)..... 53
Figura 15 - Boxplot (a) e time series plot (b) do TOG-G para o conjunto de dados balanceado (Fonte: autoria própria)..... 54
Figura 16 - Correlograma para as variáveis preditoras selecionadas e TOG-G (Fonte: autoria própria) 55
Figura 17 - Efeitos principais para a previsão do TOG-G..... 65
Figura 18 - Gráfico de linha de contorno para o modelo do TOG-G (Fonte: autoria própria) 66
Figura 19 - Gráfico de superfície para o modelo do TOG-G (Fonte: autoria própria)... 66
Figura 20 - Séries de TOG-G real e previsto para o conjunto de teste da base diária (Fonte: autoria própria) 68
Figura 21 - Séries de TOG-G real, média e 3º quartil dos TOG-G previstos para o conjunto de teste da base fiscal (Fonte: autoria própria)..... 71

Figura 22 - Séries de TOG-G real, média e 3º quartil dos TOG-G previstos (a) com agrupamento e (b) sem agrupamento para o conjunto de teste da base fiscal (Fonte: autoria própria) 76
Figura 23 - Séries de TOG-G real e TOG-S real para o conjunto de teste da base diária (Fonte: autoria própria)..... 78
Figura 24 - Matrix plot do TOG-G diário real e TOG-G diário previsto (Fonte: autoria própria) 79
Figura 25 - *Matrix plot* do TOG-G diário real e média do TOG-G fiscal previsto (Fonte: autoria própria) 79
Figura 26 - *Matrix plot* do TOG-G diário real e 3o. quartil do TOG-G fiscal previsto (Fonte: autoria própria)..... 79
Figura 27 - *Matrix plot* do TOG-G diário real e TOG-S diário real (Fonte: autoria própria) 80
Figura 28 - Boxplot do Peso (Fonte: autoria própria) 83
Figura 29 - Boxplot (a), time series plot (b) e gráfico de barras (c) do Peso para o conjunto de dados balanceado (Fonte: autoria própria)..... 84
Figura 30 - Efeitos principais para a previsão do Peso..... 89
Figura 31 - Gráfico de linha de contorno para o modelo do Peso (Fonte: autoria própria) 90
Figura 32 - Gráfico de superfície para o modelo do Peso (Fonte: autoria própria)..... 90
Figura 33 - Séries de Peso real e previsto (com agrupamento) para o conjunto de teste (Fonte: autoria própria)..... 92
Figura 34 - Séries de Peso real e previsto (sem agrupamento) para o conjunto de teste (Fonte: autoria própria)..... 94
Figura 35 - Matrix plot do Peso real e Peso previsto (com agrupamento) (Fonte: autoria própria) 95
Figura 36 - *Matrix plot* do Peso real e Peso previsto (sem agrupamento) (Fonte: autoria própria) 95

LISTA DE TABELAS

Tabela 1 - Quantidade de valores faltantes na base diária 48
Tabela 2 - Limites inferior e superior e quantidade de registros para cada agrupamento 52
Tabela 3 - ANOVA das variáveis mais significativa para a previsão do TOG-G 54
Tabela 4 - Correlação de Pearson entre as variáveis preditoras selecionadas e o TOG-G 55
Tabela 5 - Parâmetros e acurácias dos algoritmos de classificação 56
Tabela 6 - Distribuição dos agrupamentos dentro dos conjuntos de treinamento e teste 57
Tabela 7 - Classificações de agrupamento para o conjunto de teste da base diária 57
Tabela 8 - Métricas de avaliação para as classificações da base diária 58
Tabela 9 - Matriz de confusão para a previsão do Agrupamento do conjunto de testes da base diária 59
Tabela 10 - Classificações de agrupamento e probabilidade associada a cada uma das classes o conjunto de teste da base fiscal 59
Tabela 11 - Coeficientes padronizados para os termos do modelo de previsão do TOG-G 65
Tabela 12 - Valores de TOG-G previstos e métricas MAE e MAPE 66
Tabela 13 - Valores de TOG-G previstos para o conjunto de teste da base fiscal 68
Tabela 14 - Média e 3º quartil dos valores de TOG-G previstos para a base fiscal 69
Tabela 15 - Métricas para os conjuntos de teste relativos às abordagens diária e fiscal e TOG-S 71
Tabela 16 - Valores de TOG-G previstos (sem agrupamento) e métricas MAE e MAPE 72
Tabela 17 - Valores de TOG-G previstos (sem agrupamento) para o conjunto de teste da base fiscal 73
Tabela 18 - Média e 3º quartil dos valores de TOG-G previstos (sem agrupamento) para a base fiscal 74
Tabela 19 - Valores de TOG-G e TOG-S reais e métricas MAE e MAPE 77
Tabela 20 - Métricas para os conjuntos de teste relativos às abordagens diária e fiscal (com e sem agrupamento) e TOG-S 80
Tabela 21 - Correlação de Pearson entre as variáveis preditoras da base de dados clássica e a variável Peso 82

Tabela 22 - Limites inferior e superior e quantidade de registros para cada agrupamento 83
Tabela 23 - ANOVA do modelo de regressão linear do Peso 85
Tabela 24 - Parâmetros e acurácias dos algoritmos de classificação 86
Tabela 25 - Distribuição dos agrupamentos dentro dos conjuntos de treinamento e conjunto de teste 86
Tabela 26 - Classificações de agrupamento para o conjunto de teste 87
Tabela 27 - Métricas de avaliação para o classificador 87
Tabela 28 - Matriz de confusão para a previsão do Agrupamento do conjunto de testes da base 88
Tabela 29 - Coeficientes padronizados para os termos do modelo de previsão do Peso 89
Tabela 30 - Valores de Peso previstos e métricas MAE e MAPE 91
Tabela 31 - Valores de Peso real e previsto (sem agrupamento) e métricas MAE e MAPE 92
Tabela 32 - Métricas para o conjunto de teste relativos às abordagens com agrupamento e sem agrupamento 94
Tabela 33 - Correlação entre os resultados do método proposto e da regressão sem o agrupamento 96

LISTA DE QUADROS

Quadro 1 - Composição dos registros da base fiscal para o dia D-0 46
Quadro 2 - Codificação das variáveis categóricas 47
Quadro 3 - Variáveis selecionadas a partir da opinião especializada 49
Quadro 4 - Composição da base de dados clássica 81

LISTA DE ABREVIATURAS E SIGLAS

TOG	Teor de óleos e graxas
CONAMA	Conselho Nacional do Meio Ambiente
TOG-G	Teor de óleos e graxas aferido pelo método gravimétrico
FPSO	*Floating, Production, Storage and Offloading* - Unidade flutuante de produção, armazenamento e transferência
OLS	*Ordinary Least Squares* – Mínimos Quadrados Ordinários
TOG-E	Teor de óleos e graxas aferido pelo método *eracheck*
TOG-S	Teor de óleos e graxas aferido pelo método espectrofotométrico
MAE	*Mean Absolute Error* - Erro Médio Absoluto
MAPE	*Mean Absolute Percentage Error* - Erro Percentual Absoluto Médio

SUMÁRIO

1. INTRODUÇÃO 16
1.1. Contexto da pesquisa 16
1.2. Objetivos 17
1.2.1. Objetivo Geral 17
1.2.2. Objetivos Específicos 17
1.3. Contribuições esperadas 18
1.4. Delimitações da pesquisa 18
1.5. Estrutura do trabalho 18
2. FUNDAMENTAÇÃO TEÓRICA 20
2.1. Processamento primário de petróleo 22
2.2. Teor de óleos de graxas 24
2.3. Regressão linear múltipla 25
2.4. Balanceamento de dados 27
2.5. Métodos de aprendizado de máquina 28
2.5.1. *Random Forest* 28
2.5.2. *K-nearest neighbors* (KNN) 30
2.5.3. Redes Neurais Artificiais 32
2.5.4. Regressão Logística 33
2.5.5. Máquinas de vetores de suporte (*Support Vector Machine* – SVM) 35
3. MÉTODO PROPOSTO 36
3.1. Considerações iniciais 36
3.2. Método 36
3.2.1. Pré-processamento e pré-análise 36
3.2.2. Classificação e previsão 38
3.3. Caracterização da pesquisa 44

4. APLICAÇÃO DO MÉTODO 46
4.1. Conjunto de dados de Teor de Óleos e Graxas (TOG) 46
4.1.1. Considerações iniciais 46
4.1.2. Pré-processamento 46
4.1.3. Pré-análise 47
4.1.4. Etapa de Classificação - A 51
4.1.5. Etapa de Previsão - B 64
4.1.6. Confirmação 71
4.2. Conjunto clássico de dados 81
4.2.1. Considerações iniciais 81
4.2.2 Pré-processamento 81
4.2.3. Pré-análise 82
4.2.4. Etapa de Classificação - A 83
4.2.5. Etapa de Previsão - B 88
4.2.6. Confirmação 92
5. CONCLUSÕES 97
5.1. Conclusões gerais 97
5.2. Contribuições do trabalho 99
5.3. Sugestões para estudos futuros 99
REFERÊNCIAS 100
APÊNDICE A – Descrição das variáveis da base de dados original 105
APÊNDICE B – Correlação de Pearson entre as variáveis preditoras da base de dados original e a variável TOG-G 109

1. INTRODUÇÃO

1.1. Contexto da pesquisa

A água produzida descartada nos oceanos em função do processamento primário de petróleo realizado nas plataformas de extração de petróleo e gás natural contém um teor de óleos e graxas (TOG) que deve ser continuamente monitorado, almejando enquadramento em limites determinados pelo órgão ambiental regulador. Níveis diários e mensais que extrapolem os estabelecidos geram, além de prejuízos ao meio ambiente, penalidades financeiras às empresas exploradoras.

No Brasil, na Resolução n° 39379 de 8 de agosto de 2007, o Conselho Nacional do Meio Ambiente (CONAMA) estabelece as normas relacionadas ao descarte contínuo de água de processo ou de produção em plataformas marítimas de petróleo e gás natural, evidenciando o conceito de água produzida como sendo a água normalmente produzida junto com o petróleo. Em seu Art. 5º, define que o descarte de água produzida deve conter uma concentração de TOG cujos limites máximo diário e mensal sejam de, respectivamente, 42 mg/L e 29 mg/L, sendo este último calculado por média aritmética simples. No subsequente artigo, determina que o único método homologado para realização desta medição do TOG é o gravimétrico. Entretanto, ressalta que o órgão ambiental poderá acatar outros métodos de análise que apresentem correlação comprovada e estatisticamente significativa com o método oficial. Ainda no Art. 15, a Resolução ressalta que valores de TOG não enquadrados nos limites pré-determinados incorrerá em sanções previstas pela legislação vigente.

O método gravimétrico mede tanto frações dispersas quanto as dissolvidas recuperadas após acidificação, com exceção das frações leves até a temperatura de evaporação do solvente utilizado durante o processo. Dentre as limitações do método estão sua análise, que deve sempre ser realizada em terra (*onshore*) devido à utilização de balanças analíticas, seu alto custo e, mais grave, a lentidão no retorno do resultado da medição. Atualmente, a média de tempo para obtenção do valor de TOG gravimétrico (TOG-G) diário de determinada plataforma é de 20 dias após a coleta da amostra.

Daí parte o problema de pesquisa a ser explorado por este estudo, dado que as consequências desta defasagem de tempo entre a coleta e o resultado da medição podem ser sentidas financeiramente pela exploradora, que sofre penalidades previstas em lei a

cada desenquadramento (diário e mensal), mas podem, principalmente, ser identificadas no meio ambiente, visto que quando o resultado se torna acessível, a água contaminada já foi descartada em mar aberto. Isto embasa a oportunidade desta tese que objetiva desenvolver, através da aplicação de técnicas de classificação e mineração de dados, um modelo preditivo de TOG gravimétrico a partir de variáveis de processo que possa ser utilizado *offshore*. Além disso, a pesquisa busca identificar os efeitos das variáveis preditoras sobre o comportamento do TOG-G, de maneira a auxiliar na tomada de decisão pela operação da plataforma.

Variáveis de processo, bem como informações advindas de outros métodos de medição utilizados nas plataformas, mas não oficializados pelo órgão regulador, como o método espectrofotométrico, serão considerados no decorrer das análises. Além disso, pretende-se incorporar informações da variável de resposta ao conjunto de variáveis preditoras, visando produzir resultados com maior acurácia e correlação com o método homologado pelo CONAMA.

1.2. Objetivos

A partir do problema de pesquisa detalhado anteriormente, é possível delinear os objetivos geral e específicos que permeiam este trabalho.

1.2.1. Objetivo Geral

O objetivo principal deste trabalho é propor um método para previsão de valores do TOG gravimétrico a partir das variáveis do processamento primário de petróleo e das medições do TOG espectrofotométrico (TOG-S), além de uma nova preditora extraída da variável de resposta. Para tal, pretendem-se explorar técnicas de classificação e de mineração de dados, de maneira a elaborar um modelo cujos resultados alcancem correlação estatisticamente significativa com os valores reais de medição do método homologado pelo CONAMA.

1.2.2. Objetivos Específicos

Diante do objetivo principal, tem-se como objetivos secundários deste trabalho identificar os efeitos das variáveis preditoras sobre o comportamento do TOG-G, bem como fornecer resultados de previsão com valor de correlação com o TOG-G superiores às existentes com os outros métodos de aferição utilizados *offshore* até o momento, como o TOG

espectrofotométrico. Também espera-se demonstrar a viabilidade do método através de sua aplicação a um conjunto de dados clássico de regressão linear.

1.3. Contribuições esperadas

Este estudo busca contribuir à indústria petroquímica no que tange à previsibilidade do teor de óleos e graxas presente na água produzida descartada nos oceanos. Com isso, apresenta-se um método combinado de classificação e previsão do TOG quantificado pelo método gravimétrico, facilitando a tomada de decisão pela operação da plataforma, a fim de prevenir danos ambientais e evitar sanções financeiras por descumprimento de normas. O método é aplicado a um conjunto real de dados de uma plataforma de petróleo, cujos resultados comprovem a utilidade e relevância do estudo. Além disso, a pesquisa visa apresentar uma estratégia para extrair e incorporar informações da variável de resposta ao conjunto de preditores, demonstrando a contribuição desta manobra ao aumento do poder de previsão do modelo.

1.4. Delimitações da pesquisa

Os dados reais a serem utilizados neste trabalho foram obtidos por meio de parceria com uma indústria petroquímica brasileira. Assim, as informações de TOG gravimétrico, essenciais para o desenvolvimento do estudo, foram disponibilizadas apenas em base diária, o que limita o confrontamento dos resultados obtidos em base horária, exigindo a aplicação de estratégias estatísticas para a aferição dos resultados. Também os dados de TOG *eracheck* não foram contemplados dentro do período completo da base histórica, o que dificultou a sua incorporação ao estudo.

1.5. Estrutura do trabalho

Este trabalho encontra-se estruturado em 5 capítulos, dentre os quais neste, o primeiro, foram apresentadas uma sumária contextualização sobre os assuntos explorados, além dos objetivos geral e específicos, as contribuições e as delimitações da pesquisa.

O segundo capítulo apresenta a fundamentação teórica associada ao tema principal do trabalho e às técnicas aplicadas para desenvolvimento do método proposto, de maneira que são detalhados o processamento primário de petróleo, bem como o conceito de teor de óleos de graxas. Também são abordadas as técnicas de regressão linear múltipla, de balanceamento de dados e os métodos de classificação aplicados.

O método proposto é detalhado no capítulo 3, que apresenta suas necessidades de pré-análise e condições de pré-processamento, as etapas de classificação e previsão nele envolvidos, e a fase de confirmação. Subsequentemente, o quarto capítulo apresenta a aplicação do método proposto a um conjunto de dados reais da indústria petroquímica e também a um conjunto clássico de dados de regressão linear.

Finalmente, o último capítulo discute as conclusões obtidas com esta pesquisa, bem como reafirma suas contribuições. Sugestões para trabalhos futuros são também mencionados neste capítulo. Os apêndices encerram o documento, trazendo informações complementares para o completo entendimento do estudo desenvolvido.

2. FUNDAMENTAÇÃO TEÓRICA

O processamento primário de petróleo realizado em plataformas *offshore* como as FPSO (*Floating, Production, Storage and Offloading*) tem como principal função a separação do petróleo bruto dos poços explorados em água, óleo e gás (DE OLIVEIRA *et al.*, 2019). A água produzida nesta separação, que contém um certo teor de óleos e graxas, pode ser reinjetada nos poços ou descartada em mar aberto (DE OLIVEIRA *et al.*, 2019; TRIGGIA *et al.*, 2001). No caso de descarte, os valores de TOG são especificados de forma a atender aos requisitos estabelecidos na resolução nº 393/2007 do CONAMA (CONAMA, 2007). A média mensal de 29mg/L e o valor máximo de 42mg/L diário são os limites indicados na resolução.

De acordo com Costa *et al.* (2022), as águas produzidas despontam como um dos principais efluentes das plataformas de óleo e gás, o que reitera a importância do aprimoramento constante dos métodos para seu adequado monitoramento, uma vez que tratamentos usuais como tanque, flotação e hidrociclones podem não remover de forma eficiente as impurezas orgânicas (STANDARD METHODS). O estudo desenvolvido por Biazon *et al.* (2019), onde foi avaliada a contribuição da incerteza da recuperação do método gravimétrico na análise dos níveis de TOG na água produzida, destaca os diversos danos ambientais e prejuízos humanos causados por um descarte em desacordo com a legislação.

Lv *et al.* (2017), por sua vez, abordaram a relevância desse assunto ao citar que um descarte inadequado da água produzida pode poluir não só a superfície, mas também a água subterrânea e o solo. Também Klemz *et al.* (2021) em seu estudo sobre o tratamento da água produzida, destacaram que um grande desafio envolvido é a presença de contaminantes, além dos compostos orgânicos dissolvidos, na água produzida.

Lee *et al.* (2011) afirmaram que diferentes métodos são usados para medir o valor de TOG na água produzida, baseados em princípios de cromatografia, gravimetria e absorção de infravermelho. No entanto, cada um deles mede uma determinada fração da amostra analisada. Para fins regulatórios, o CONAMA determina como método homologado o gravimétrico ou qualquer outra metodologia significativamente correlacionada a ele (CONAMA, 2007).

Um dos grandes problemas enfrentados pelas plataformas FPSO, objeto de estudo deste trabalho, é o intervalo de tempo de cerca de 20 dias para acessar a medição oficial do TOG-G diário. Isto ocorre devido à necessidade dessa análise ser realizada *onshore*, o que exige esforços logísticos para a transferência da amostra para o laboratório. Essa defasagem entre a coleta e a resposta dificulta a tomada de decisão sobre ações preventivas ou corretivas para manter o TOG-G diário e, consequentemente, mensal dentro dos limites regulatórios.

Diante deste cenário, e considerando que o Brasil continua entre os principais produtores de petróleo do mundo, de acordo com o último relatório *Statistical Review of World Energy* publicado pela *British Petroleum* (COMPANY, 2021), torna-se eminente avaliar a viabilidade de construção de um modelo para prever os valores de TOG-G na plataforma *offshore* a partir de informações sobre variáveis de processo e também de medições de TOG geradas por outras metodologias. O acesso à base histórica de dados pode auxiliar na identificação de faixas de TOG-G que venham a produzir informações relevantes que possam ser incorporadas ao modelo, contribuindo para uma maior acurácia em relação à previsão de não conformidade.

Diversos são os contextos de aplicação de técnicas de regressão e classificação. Akinbinu (2010) investigou a existência de correlação entre propriedades como gradiente de fratura, pressão de sobrecarga, pressão de poro e profundidade de campo vertical verdadeira e, a partir daí, desenvolveu modelos para estimar o gradiente de fratura por meio de técnicas estatísticas de regressão múltipla com *stepwise*. Davtyan *et al.* (2020) desenvolveram um modelo de regressão dinâmica para prever as taxas de produção de petróleo em um campo de exploração real, por meio do uso de aprendizado de máquina.

Devido à importância da característica de permeabilidade de rochas naturalmente fraturadas ao simular o fluxo no reservatório de hidrocarbonetos, Habibi *et al.* (2014) propuseram uma equação prática consistindo em distintas características fractais e estatísticas de padrões fraturados por meio de análise de regressão multivariada. Sleiti *et al.* (2021), por sua vez, aplicaram técnicas de regressão múltipla não linear para construir uma nova correlação universal para a razão gás-óleo, uma vez que a acurácia dos valores previstos para as propriedades de pressão-volume-temperatura são cruciais para modelagem e simulação confiáveis de detecção de *kick* de gás.

Ainda tratando de questões práticas na operação de poços e linhas de petróleo e gás, Marins *et al.* (2021) utilizaram o algoritmo *Random Forest* para construir um classificador de falhas a ser implementado em um sistema de detecção automática e classificação de eventos com falhas reais. Hegde *et al.* (2019), com o objetivo de alcançar um modelo baseado em dados para minimizar as vibrações da coluna de perfuração, utilizaram algoritmos de classificação de aprendizado de máquina, como *Support Vector Machine* (SVM), Regressão Logística (RL), *Random Forest* (RF), dentre outros, para classificar como baixo ou alto o índice *stick slip*, identificando que o *Random Forest* atingiu uma precisão média de 90%.

No presente estudo, uma combinação de métodos de classificação e regressão foi aplicada para melhorar a acurácia da previsão dos valores de TOG-G. As medições de TOG-G foram divididas em quatro faixas, às quais foram associados agrupamentos, e um modelo matemático de previsão de TOG-G foi construído para cada um deles. A fim de direcionar corretamente novas observações ao agrupamento adequado, um modelo de classificação também foi construído. A utilização da regressão linear múltipla foi proposta devido à facilidade de implementação e interpretação de seus resultados, enquanto que o algoritmo de *Random Forest* foi escolhido por ter apresentado melhores resultados para a classificação dos agrupamentos.

2.1. Processamento primário de petróleo

O petróleo é formado a partir de uma mistura de hidrocarbonetos com uma variedade de componentes químicos geralmente insolúveis em água, cuja separação em componentes puros ou misturas de composição conhecida é um processo extremamente complexo (DE OLIVEIRA *et al.*, 2019). Dessa forma, costuma ser separado em frações de acordo com a faixa de ebulição de seus compostos: gás residual, gás liquefeito de petróleo (GLP), gasolina, querosene, diesel leve, diesel pesado, lubrificantes e resíduos (TRIGGIA *et al.*, 2001).

A plataforma FPSO é um tipo de unidade de processamento primário de petróleo do tipo flutuante, como pode ser observado na Figura 1, capaz de produzir, armazenar e transferir petróleo e gás (HAN; ZHEN; HUANG, 2021). Ela dispõe de alta tecnologia e é usualmente aplicável a cenários de exploração de campos de petróleo em águas profundas ou ultraprofundas, cuja infraestrutura de dutos está ausente ou prejudicada (TRIGGIA *et*

al., 2001). O óleo produzido pela FPSO é escoado por navios-aliviadores, enquanto que o escoamento de gás é realizado por meio de dutos submarinos.

Figura 1 - Plataforma FPSO (Fonte: (Agência Brasil, 2017))

Uma de suas vantagens é que toda a sua estrutura pode ser reaproveitada por meio de sua migração para outro ponto tão logo se esgote o óleo local (ALLAHYARZADEH-BIDGOLI *et al.*, 2018; TRIGGIA *et al.*, 2001). Critérios como as condições do mar, a distância da superfície da água ao fundo do mar, a finalidade do poço, a logística de apoio, entre outros, determinam a instalação deste tipo de plataforma. Embora existam sistemas de ancoragem e posicionamento dinâmico que garantem que a plataforma permaneça em um determinado local na superfície do mar, os sistemas de medição alocados a ela invariavelmente sofrem interferências devido à ação constante de ondas, correntes e ventos sobre a plataforma (ALLAHYARZADEH-BIDGOLI *et al.*, 2018).

Dentre as atividades realizadas na FPSO, tem-se o tratamento da água produzida, que visa recuperar uma parte do óleo ainda presente na emulsão e proporcionar as condições necessárias para a reinjeção ou descarte da água, dentro de condições pré-definidas de preservação ambiental (LEE; NEFF, 2011; TRIGGIA *et al.*, 2001). Segundo Lee e Neff (2011), a água produzida possui uma composição orgânica e inorgânica complexa que inclui, entre muitas outras, frações dissolvidas e dispersas de óleos e graxas. À medida que o poço se aproxima do final de sua vida útil, a produção de petróleo bruto diminui e a extração de água produzida associada aumenta (ALLAHYARZADEH-BIDGOLI *et al.*, 2018).

Dentre os processos de separação óleo/água bastante utilizados na indústria petrolífera, tanto em ambientes *onshore* como *offshore*, estão os tanques separadores, os hidrociclones e os flotadores (TRIGGIA *et al.*, 2001). Eventualmente, combinações de

processos, como o uso de hidrociclones seguido de flotador, são aplicadas para o tratamento da água produzida, como pode ser visualizado na Figura 2.

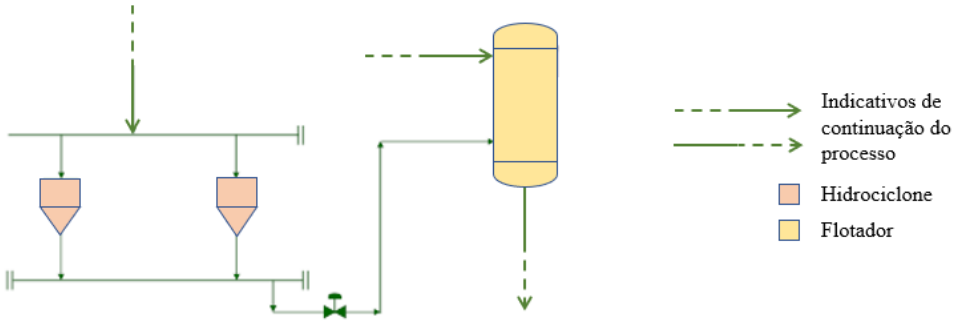


Figura 2 - Exemplo de posicionamento de hidrociclones e flotador em uma parte do processo de tratamento de água de uma FPSO (Fonte: autoria própria)

2.2. Teor de óleos de graxas

Segundo a organização *Standard Methods*, óleos e graxas podem ser definidos como qualquer material recuperado como uma substância solúvel no solvente, já que a determinação da sua concentração não parte da medição da quantidade absoluta de uma substância especificada, mas, sim, substâncias semelhantes em termos de composição são aferidas com base em sua solubilidade comum em um solvente orgânico.

O óleo na água produzida é composto por uma fase dispersa, que contém hidrocarbonetos, asfaltenos, resinas, entre outros, e uma fase dissolvida, composta por ácidos naftênicos, aromáticos, fenólicos, etc. (LEE; NEFF, 2011). Como previamente mencionado, a água produzida descartada pelas plataformas, ou seja, que está associada à produção de petróleo, tem um valor de TOG-G cuja concentração máxima deve obedecer aos órgãos reguladores, por se tratar de poluente ambiental (ARAUJO FILHO *et al.*, 2020; TRIGGIA *et al.*, 2001).

De acordo com Fan *et al.* (2018), existem diferentes métodos de quantificação de TOG que podem ser realizados *onshore* ou *offshore*, a maioria deles baseados em princípios de cromatografia, gravimetria e absorção de infravermelho. Uma vez que a concentração de TOG é estritamente dependente do método utilizado, o que significa que diferentes métodos podem gerar diferentes valores de medição, o CONAMA determina como homologado o método gravimétrico ou outra metodologia que apresente correlação estatisticamente significativa com ele (CONAMA, 2007; LEE; NEFF, 2011).

Os métodos gravimétricos podem ser aplicados a qualquer tipo de óleo e precisam ser realizados em laboratórios *onshore*, pois requerem a utilização de balanças analíticas que

sofreriam interferência da instabilidade dos ambientes *offshore*, causando perda de precisão na análise das amostras coletadas. Por meio destes métodos é possível medir tanto frações dispersas quanto dissolvidas recuperadas após acidificação, exceto frações leves até a temperatura de evaporação do solvente (LEE; NEFF, 2011). O método gravimétrico SM 5520B, aplicado na plataforma objeto do estudo de caso desta pesquisa, constitui um método de extração líquido-líquido aplicável quando objetiva-se determinar substâncias extraíveis por n-hexano, não adsorvíveis em sílica-gel, e possui um limite de quantificação de 5mg/L (Petrobras, 2019).

Devido à logística de transporte entre a FPSO e o laboratório, os resultados da medição TOG-G só são acessíveis aproximadamente 20 dias após o envio para terra firme. Isto significa que, no momento em que o operador tem acesso à medida oficial de TOG-G presente na água produzida, ela pode já ter sido descartada, causando danos ambientais e prejuízos financeiros previstos em lei (CONAMA, 2007).

Para lidar alternativamente com essa defasagem entre o dia da coleta e o resultado da medição, são utilizados instrumentos de bancada, a fim de orientar medidas corretivas imediatas. Dentre as técnicas utilizadas, são aplicadas a colorimétrica, neste caso a espectrofotometria, e a infravermelha (*eracheck*).

O método espectrofotométrico aplica-se quando da determinação de substâncias solúveis em n-hexano que possam ser detectadas através de espectrofotometria de absorção molecular em comprimentos de onda de 400nm (Petrobras, 2019b). Esta técnica tende a indicar valores inferiores às medidas realizadas pelo gravimétrico e *eracheck*, pois contempla apenas a fração dispersa. Seu limite de quantificação é de 2mg/L.

Já a medição do *eracheck* é realizada por meio de analisador que permite determinar materiais que possam ser extraídos com solvente ciclo-hexano por espectroscopia de infravermelho não dispersivo (Petrobras, 2020). Ainda que amplamente utilizado em todo o mundo e capaz de medir ambas as fases (dispersa e dissolvida), esta técnica tende a emitir valores maiores quando comparado aos outros dois métodos em questão, já que, diferente do TOG gravimétrico, ele também quantifica as frações leves até temperatura de evaporação do solvente por ela utilizado (LEE; NEFF, 2011).

2.3. Regressão linear múltipla

A metodologia estatística denominada análise de regressão é uma das técnicas de dependência mais amplamente difundidas, pois é aplicável a cenários onde objetivos de predição e/ou de explicabilidade, são perseguidos (HAIR JR. *et al.*, 2019). O primeiro diz

respeito a quanto as variáveis preditoras (x) são capazes de, de fato, prever a variável dependente (y), enquanto que o segundo explora os coeficientes da regressão em termos de sinal, magnitude e significância estatística, avaliando o efeito e a importância de cada x sobre o y (HAIR JR. *et al.*, 2019).

A análise de regressão múltipla considera a relação linear estatística existente entre uma ou mais variáveis preditoras e uma variável dependente (EVERITT; RENCHER, 1996; HAIR JR. *et al.*, 2019). Embora todas as variáveis envolvidas na regressão devam ser contínuas, eventualmente informações não métricas (ordinais ou nominais) precisam ser incorporadas na análise. Para tal, a codificação apoiada por variáveis *dummy*, por exemplo, deve ser aplicada (HAIR JR. *et al.*, 2019; SHARMA, 1996). Partindo de um modelo matemático e considerando determinado conjunto de variáveis independentes (x), conforme a Eq. (1), é possível prever o valor da variável resposta (y), bem como depreender os efeitos das variáveis preditoras sobre ela (JOHNSON; WICHERN, 2007).

$$y = \beta_0 + \beta_1 x_1 + \ldots + \beta_n x_n + \varepsilon \qquad (1)$$

onde β's são os coeficientes da regressão, n é o número de x's e ε é um erro randômico.

De maneira a se obter a melhor previsão a partir do conjunto de preditoras, cada variável x é ponderada durante a análise de regressão e este peso indica sua contribuição relativa e, consequentemente, sua influência no processo de previsão (HAIR JR. *et al.*, 2019). É importante ressaltar que o poder de predição individual de uma variável independente sofre impacto caso haja multicolinearidade no conjunto de variáveis x. Neste caso, a variância única explicada por cada preditora reduz à medida que a colinearidade aumenta, já que o percentual de predição compartilhada só pode ser contabilizado uma vez (HAIR JR. *et al.*, 2019).

Para estimar os coeficientes $\boldsymbol{\beta}$ da regressão, o método dos mínimos quadrados ordinários (*Ordinary Least Squares* - OLS) é considerado o melhor estimador linear imparcial (MARDIA; KENT; BIBBY, 1995). Sua aplicação parte de algumas premissas em torno da Eq. 1, tal como detalhado nas Eq. (2).

$$E(\varepsilon) = 0, \qquad V(\varepsilon) = \Omega \qquad (2)$$

Além disso, se ε não é normalmente distribuído, então é possível obter uma aproximação dos coeficientes $\boldsymbol{\beta}$ pelo método OLS, conforme Eq. (3):

$$\widehat{\boldsymbol{\beta}} = (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'y \tag{3}$$

Assim, o OLS busca, para determinados valores de x, minimizar a soma dos quadrados da diferença entre a resposta observada $Y_{i_{obs}}$ e o valor estimado $\hat{Y}_{i_{est}}$, de maneira que ambos valores se aproximem tanto quanto possível (EVERITT; RENCHER, 1996; JOHNSON; WICHERN, 2007), conforme apresentado na Eq. (4).

$$\begin{aligned} S(x) &= \sum_{j=1}^{w} \left(y_j - \beta_0 - \beta_1 x_{j1} - \ldots - \beta_n x_{jn}\right)^2 \\ &= (y - \boldsymbol{X}\boldsymbol{\beta})'(y - \boldsymbol{X}\boldsymbol{\beta}) \end{aligned} \tag{4}$$

Nos problemas onde a previsão é definida como principal objetivo, é imprescindível que a acurácia do modelo seja avaliada através de métricas e testes estatísticos, por meio dos quais é possível averiguar a significância do poder de predição (HAIR JR. *et al.*, 2019).

2.4. Balanceamento de dados

A construção de modelos de classificação preditiva mediante a aplicação de técnicas de aprendizado de máquina (*machine learning*) e mineração de dados (*data mining*) exige um pré-processamento criterioso que garanta a consistência do conjunto de dados e o balanceamento dos registros que o compõem, em função da classe determinada (LIN *et al.*, 2017). Isto porque a maioria dos algoritmos de classificação assumem uma quantidade similar de registros na base de treinamento para cada uma das classes identificadas. O desbalanceamento de uma base é, portanto, identificado quando uma das classes está associada a um número de registros consideravelmente menor quando comparado às outras. Este desbalanceamento pode ser sutil ou severo e é especialmente crítico quando a classe de interesse é a minoritária, visto que o modelo encontra dificuldades em reconhecer seu padrão de comportamento, aumentando as chances de distorção na classificação (BROWNLEE, 2020).

Dentre as técnicas disponíveis para lidar com os problemas de balanceamento, o *undersampling* envolve a redução da base de dados original por meio da remoção aleatória de registros associados às classes majoritárias (YAP *et al.*, 2014). A partir da

aplicação deste algoritmo, amplamente utilizado por ser de fácil implementação, a base resultante manterá todos registros da classe minoritária ao mesmo tempo em que garante uma equalização na distribuição das classes.

2.5. Métodos de aprendizado de máquina

Nesta subseção serão apresentados os métodos de aprendizado de máquina aplicados nesta pesquisa. Vale destacar que tratam-se de métodos que podem ser utilizados tanto em problemas de regressão como de classificação. Entretanto, o enfoque será mantido apenas aos problemas de classificação, uma vez que no método proposto, estes algoritmos foram utilizados para a classificação das observações quanto aos agrupamentos aos quais elas estão associadas, e não para a previsão final do TOG-G.

2.5.1. *Random Forest*

O algoritmo *Random Forest*, um dos mais difundidos métodos de aprendizado de máquina supervisionados, baseia-se nos conceitos de árvores de decisão (*Decision Trees*) e *Bagging (Bootstrap aggregation*) (BREIMAN, 2001). O algoritmo pode ser aplicado a contextos em que a variável de resposta seja categórica ou contínua, sendo então caracterizado como um algoritmo de classificação ou de regressão, respectivamente (ZHANG; MA, 2012). O aprendizado de máquina supervisionado é um processo indutivo onde as classes às quais os objetos pertencem são conhecidas e a coleção de regras que leva a esta classificação é aprendida a partir de exemplos ou instâncias contidos em um conjunto de treinamento (DOUGHERTY, 2013).

Decision Trees são algoritmos que consideram uma estrutura de árvore, com ramificação direcionada, para o procedimento de decisão hierárquico em relação à classificação (LEE; ULLAH; WANG, 2020). O seu esquema, exemplificado na Figura 3, se inicia em um nó raiz que se subdivide em dois ramos que acessam nós de decisão a partir do qual sucessivas subdivisões semelhantes podem ser realizadas até que se alcancem os nós terminais, também conhecidos como folhas, aos quais as classes são atribuídas (ROKACH, 2010). Depois de escolhidas as características determinantes para a classificação dentro do conjunto de treinamento, existe uma infinidade de possibilidades para a estruturação da árvore de decisão (DOUGHERTY, 2013).

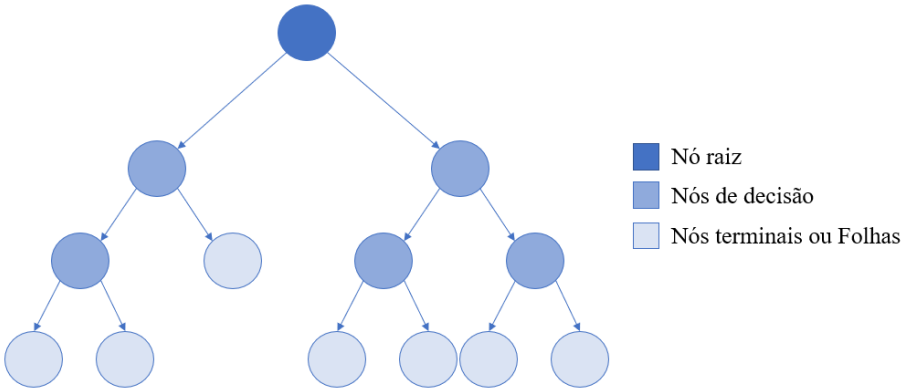


Figura 3 – Estrutura de uma árvore de decisão de 2 níveis (Fonte: autoria própria)

O conceito principal de *Bagging*, segundo Breiman (1996) e Lee *et al.* (2020), está relacionado à obtenção da média de uma coleção de previsões geradas por modelos treinados a partir de diferentes amostras extraídas de um único conjunto de dados, de maneira a obter um resultado melhor que o alcançado com a previsão a partir de um único conjunto de treinamento, reduzindo o erro de previsão. Para problemas de classificação, a classe é determinada a partir daquela com maior ocorrência dentre os resultados obtidos pelas diferentes amostras.

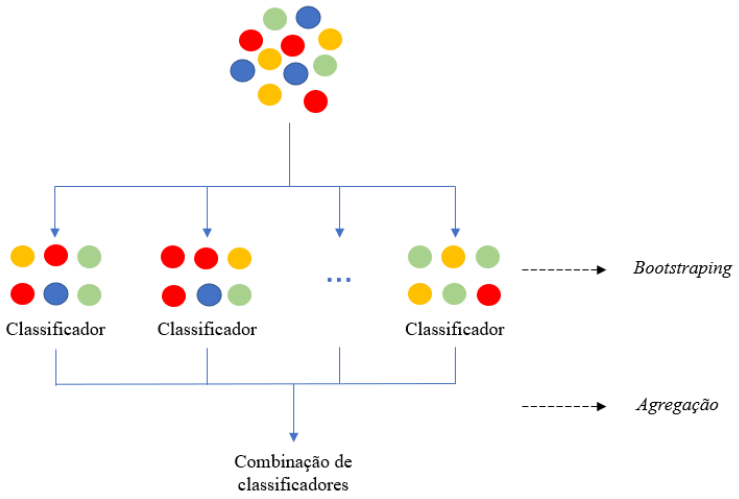


Figura 4 - Ilustração do conceito de bagging (Fonte: adaptado de Sen, 2021)

O *Random Forest*, portanto, constitui numa combinação de árvores classificadoras não correlacionadas, onde cada árvore relaciona-se a valores de um conjunto de variáveis aleatório e com a mesma distribuição (BREIMAN, 2001; HASTIE; TIBSHIRANI; FRIEDMAN, 2017; MÜLLER; GUIDO, 2017; ZHANG; MA, 2012). Para a tomada de

decisão, o algoritmo provê uma predição a partir da qual são geradas probabilidades para cada classe possível. A decisão final é alcançada quando a classe de maior probabilidade é encontrada, dentre as médias de probabilidades indicadas por todas as árvores para cada uma das classes (HASTIE, 2009; MÜLLER; GUIDO, 2017).

Dois parâmetros denominados número de estimadores ($n_{estimators}$) e número máximo de características ($max_{features}$) devem ser considerados na composição da *Random Forest*. O primeiro refere-se ao número de árvores que comporá a floresta randômica e, quanto maior, melhor será o resultado da classificação. O segundo está relacionado à quantidade de características que será selecionada para compor randomicamente a divisão associada a cada árvore. Uma *rule of thumb* para definir o parâmetro é $max_{features} = \sqrt{n_{features}}$, onde $n_{features}$ significa o número de características envolvido no problema (MÜLLER; GUIDO, 2017).

O *Random Forest* também permite mensurar a importância de cada um dos preditores utilizados para a classificação, a partir da relevância associada às características nas árvores individualmente. Quanto maior a importância, maior é a contribuição da variável para o processo de decisão.

2.5.2. *K-nearest neighbors* (KNN)

O método denominado *K-nearest neighbors* não gera um modelo preditivo, mas, sim, utiliza toda a informação contida no conjunto de dados a cada nova previsão (SANQUETTA et al., 2018; ZHANG et al., 2018a; ZUO; ZHANG; WANG, 2008). O princípio do método baseia-se na definição de um valor para o parâmetro *k* que indica a quantidade de pontos a serem considerados. Esse valor é escolhido empiricamente pelo pesquisador, embora existam diversos trabalhos disponíveis na literatura que auxiliem essa escolha (HASSANAT et al., 2014; ZHANG et al., 2018b).

Após a determinação de *k*, calcula-se a distância de uma nova observação a cada um dos pontos pertencentes ao conjunto de dados. Dentre as métricas de distância, as mais comumente utilizada são: Euclidiana, Minkowski e Manhattan. As fórmulas para cálculo das duas primeiras estão detalhadas nas Eq. (5) e Eq. (6), de acordo com El-Dahshan e Bassiouni (2018), respectivamente. Já a distância de Manhattan, possui fórmula equivalente à Eq. (6) quando $q = 1$ (ZHANG, 2012). É importante ressaltar que $\boldsymbol{X}$ e $\boldsymbol{Y}$ representam dois vetores compostos por um total de *p* componentes.

$$d(\boldsymbol{X}, \boldsymbol{Y}) = \sum_{k=1}^{p} \|X_k - Y_k\|^2 \tag{5}$$

$$d(\boldsymbol{X}, \boldsymbol{Y}) = \left(\sum_{k=1}^{p} |x_k - y_k|^q \right)^{1/q} \tag{6}$$

Considerando um contexto de classificação, a classe real associada aos dados do conjunto de teste deve ser fornecida, o que caracteriza o problema como sendo um algoritmo de aprendizado de máquina supervisionado (UDDIN et al., 2022).

A partir do cálculo das distâncias, selecionam-se os *k* vizinhos mais próximos da nova observação. A classe prevista para a nova observação é, então, definida por meio da classe mais frequente dentro do conjunto dos vizinhos mais próximos.

A Figura 5, adaptada de Müller e Guido (2016), mostra a utilização do algoritmo KNN para um conjunto de dados com duas classes distintas representadas por pentágonos roxos e trapézios verdes. Para esse caso tem-se $k = 3$. Dessa forma, as novas observações, representadas pelos círculos, são classificadas com a mesma cor da classe predominante entre os 3 vizinhos mais próximos.

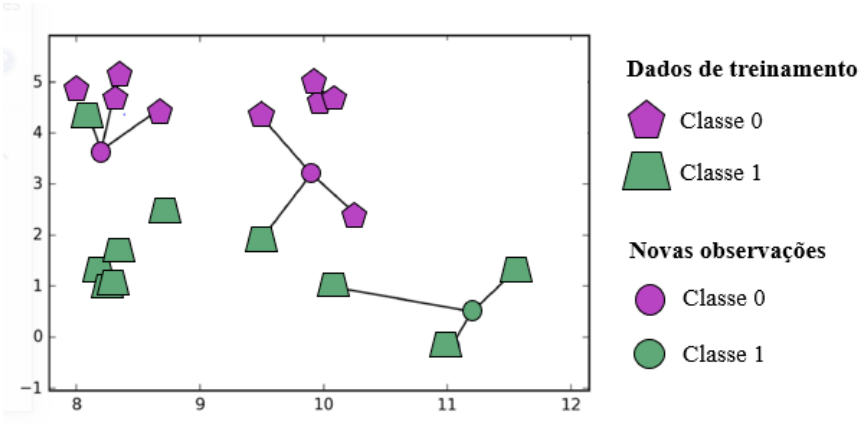


Figura 5 - Exemplo de classificação via KNN (Fonte: adaptado de Müller e Guido (2016))

2.5.3. Redes Neurais Artificiais

As redes neurais artificiais (*Artificial Neural Network* - ANN) constituem uma abordagem comumente utilizada em problemas de classificação, regressão, séries temporais e análise de cluster. As ANN são formadas por neurônios agrupados em camadas, sendo uma camada de entrada, uma ou mais camadas escondidas e uma camada de saída. A Figura 6 apresenta o exemplo de uma ANN com uma camada escondida. Neurônios de uma mesma camada não possuem ligação entre si, ao passo que estão conectados com todos os neurônios das camadas adjacentes (BALESTRASSI et al., 2009).

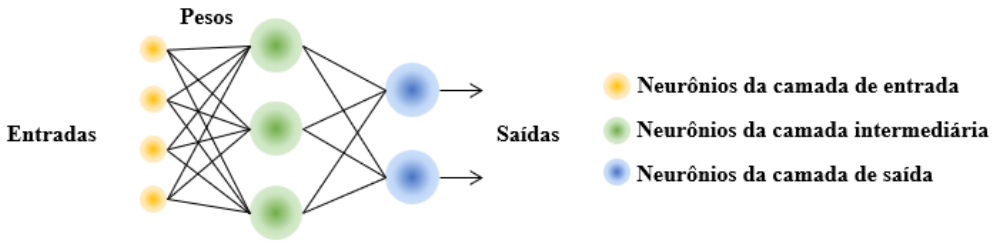


Figura 6 - Exemplo de uma ANN com uma camada escondida (Fonte: adaptado de Balestrassi et al., 2009)

A rede denominada *Multi-Layer Perceptron* (MLP) é amplamente aplicada a problemas de classificação, como pode ser observado em (CASTELLANI, 2018; CHAU; THOAI; DAO, 2021; LAZRI; AMEUR, 2018; LIN et al., 2021; WANG; MOAYEDI; FOONG, 2020). Seus principais parâmetros envolvem a quantidade de camadas escondidas, a quantidade de neurônios contidos nessas camadas e a função de ativação utilizada. Ainda, as camadas de uma rede podem estar dispostas de uma maneira *feedback* ou *feedforward* onde, naquela, uma mesma unidade de processamento pode ser visitada mais de uma vez (YEGNANARAYANA, 2005).

Usualmente, o algoritmo *backpropagation* é aplicado para o treinamento das ANNs (CHEN et al., 2018). Por se tratar de uma forma de aprendizado supervisionado, este algoritmo é frequentemente utilizado de forma a ajustar os pesos que ligam um neurônio de uma camada aos neurônios das camadas subsequentes. Inicialmente, em uma rede do tipo *feedforward* treinada pelo algoritmo *backpropagation*, a informação flui da camada de entrada para a de saída e em seguida a resposta produzida é comparada com a resposta real e, então, ajustam-se os pesos a fim de reduzir os erros (SILVA et al., 2016).

Uma grande vantagem no uso das redes neurais é que elas não necessitam de modelo prévio a respeito do conjunto de dados para conseguirem mapear as relações existentes entre os conjuntos de dados de entrada e saída (HAYKIN, 2009).

2.5.4. Regressão Logística

A regressão logística é um modelo estatístico capaz de estimar a probabilidade associada à ocorrência de determinado evento a partir de um conjunto de variáveis explicativas (HOSMER JR.; LEMESHOW; STURDIVANT, 2013). Como pode ser observado na Figura 7, para problemas de classificação contendo apenas duas classes possíveis como resposta, utiliza-se a regressão logística binária. Entretanto, quando as observações podem ser alocadas em 3 ou mais classes distintas, aplicam-se as técnicas de regressão logística denominadas multinominal ou ordinal.

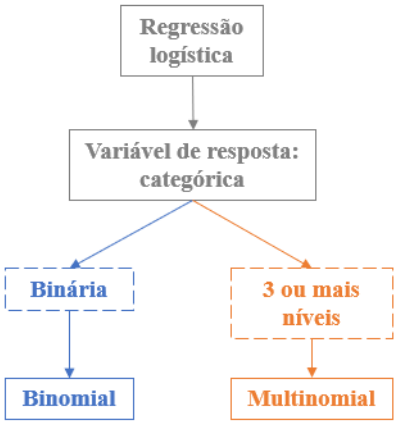


Figura 7 - Tipos de regressão logística (Fonte: autoria própria)

Um problema envolvendo variáveis nominais com mais de dois níveis pode ser exemplificado através da determinação da cor dos olhos de bebês: azul, verde, castanho ou preto. Não existe uma ordenação estabelecida para as categorias, assim como seria o caso de classificar pacientes em sadio ou doente, fumante ou não fumante. Entretanto, existem cenários em que existe uma ordem pré-estabelecida para as categorias, o que caracteriza as variáveis como sendo ordinais. Resultados ordinais podem incluir variáveis como extensão de determinada doença (nenhuma, alguma, grave), desempenho de um colaborador em uma empresa (inadequado, satisfatório, excelente) e respostas em uma pesquisa de opinião (discordo totalmente, discordo, concordo, concordo totalmente) (HOSMER JR.; LEMESHOW; STURDIVANT, 2013).

2.5.4.1. Regressão logística multinomial

A regressão logística multinomial é uma técnica utilizada quando a variável dependente é nominal. Considerando um caso em que 3 classes são apresentadas (*0*, *1* e *2*) e um vetor com *p* variáveis preditoras, podem-se, então, estimar as duas funções logit, de acordo com Hosmer Jr., Lemeshow e Sturdivant (2013), conforme mostrado na Eq. (7), onde $i = 1, 2$.

Disso, é possível obter as probabilidades de pertencimento a cada classe por meio da expressão genérica para um modelo com 3 categorias conforme apresentado em Hosmer Jr., Lemeshow e Sturdivant (2013), apresentado na Eq. (8).

$$\begin{aligned} g_i(\boldsymbol{x}) &= \ln\left[\frac{\mathrm{P}(Y = i|\boldsymbol{x})}{P(Y = 0|\boldsymbol{x})}\right] \\ &= \beta_{i0} + \beta_{i1}x_1 + \beta_{i2}x_2 + \beta_{ip}x_p \\ &= \boldsymbol{x}'\boldsymbol{\beta_i} \end{aligned} \tag{7}$$

$$\pi_j(\boldsymbol{x}) = P(Y = j|\boldsymbol{x}) = \frac{e^{g_j(\boldsymbol{x})}}{\sum_{k=0}^{2} e^{g_k(\boldsymbol{x})}} \tag{8}$$

2.5.4.2. Regressão logística ordinal

A regressão logística ordinal, segundo Hosmer Jr., Lemeshow e Sturdivant (2013), é aplicada ao classificar conjuntos com 3 ou mais classes ordinais. Inicialmente, um modelo linear é calculado e então são obtidos os valores dos coeficientes $\boldsymbol{\beta}$ da equação da função de ligação escolhida, neste caso a logit, dada pela Eq. (9).

$$g(\chi) = \log_e\left(\frac{\chi}{1 - \chi}\right) \tag{9}$$

O cálculo da probabilidade de pertencimento da observação a cada uma das classes é realizado segundo a Eq. (10), onde tem-se a constante θ_k, com $k = 1, \ldots, m - 1$ e m representando o número de classes, e o valor da multiplicação $\boldsymbol{X\beta}$ fornecendo uma matriz com n linhas e $m - 1$ colunas, sendo n o número total de observações do conjunto.

$$P(y \leq k) = e^{\left(\frac{\theta_k + \boldsymbol{X\beta}}{1 + e^{(\theta_k + X\beta)}}\right)} \tag{10}$$

2.5.5. Máquinas de vetores de suporte (*Support Vector Machine* – SVM)

As máquinas de vetores de suporte baseiam-se em construir, a partir dos dados de treinamento, uma superfície capaz de separar as diferentes classes, aprendendo a importância de cada observação para definir esse limite. Este algoritmo é capaz de separar dados inicialmente não separáveis linearmente, adotando as chamadas funções de *kernel*, que possibilita a separação linear dos dados, como pode ser observado na Figura 8 (SIMÕES et al., 2021).

A estratégia de aumento de dimensionalidade utilizada pelo SVM é essencial em termos de separação das classes. Entretanto, isto aumenta consideravelmente o tempo de processamento. Para lidar com tal questão, usa-se uma manobra matemática conhecida como '*the kernel trick*', de maneira que as funções do *kernel* não realizem as transformações de dimensões de fato, mas apenas calculem as relações entre cada par de pontos como se estivessem nas dimensões superiores (MÜLLER; GUIDO, 2016).

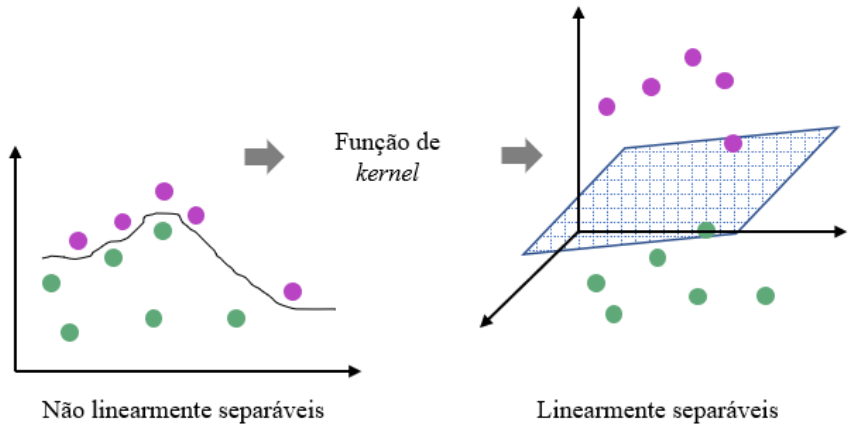


Figura 8 - Exemplo de classificação via SVM (Fonte: adaptado de Müller e Guido (2016))

Ainda segundo Müller e Guido (2016), inicialmente são selecionados hiperplanos que apresentem o maior número de acertos em termos de classificação. Porém, alguns erros de classificação no conjunto de treinamento são permitidos para evitar uma baixa performance no conjunto de teste. Na sequência, busca-se maximizar a distância entre o hiperplano e as classes consideradas. Finalmente, a região delimitada pelo hiperplano determinará à qual classe deve estar associada a nova observação. A quantidade de erros de classificação permitidos e até qual observação terá influência sobre a construção do hiperplano são parâmetros definidos pelo pesquisador, visando maximizar o número de acertos do modelo.

3. MÉTODO PROPOSTO

3.1. Considerações iniciais

O presente trabalho tem como objetivo desenvolver um método de previsão horária do TOG-G a partir das variáveis do processamento primário de petróleo e de medidas de TOG aferidas por outros métodos, para utilização em plataformas *offshore* em tempo real, uma vez que a medida gravimétrica oficial obtida em terra é usualmente divulgada com certa defasagem de tempo. Para aumentar a capacidade preditiva do modelo, pretende-se incorporar informações relativas à variável de interesse ao conjunto de variável preditoras. A alta disponibilidade da previsão do TOG-G é justificada para que o operador da plataforma atue de maneira mais ágil sobre a mitigação do risco de não conformidade do TOG-G, o que evitaria penalidades financeiras e, principalmente, prejuízos ambientais.

3.2. Método

3.2.1. Pré-processamento e pré-análise

Um dos grandes desafios do desenvolvimento de um modelo de previsão horária para o TOG-G é o fato de ser realizada apenas uma medição por dia, ainda que a amostra de água produzida enviada para análise seja composta por 4 subamostras misturadas, coletadas em 4 horários distintos e pré-determinados ao longo do dia. Os horários em que as amostras são coletadas são conhecidos como horários fiscais.

Usualmente, a base histórica da plataforma armazena dados 24 horas por dia. Assim, julgou-se adequado determinar a extração de apenas os registros relacionados aos 4 períodos fiscais de cada dia para a compilação de novas bases para análise, visto que representam os momentos em que as informações seriam mais relevantes. Desta forma, propõe-se que seja construído um banco de dados denominado base fiscal a partir da base de dados histórica. Ele visa contemplar informações relativas aos horários de coleta das subamostras, ou seja, armazenar 4 registros para cada dia de medição do TOG-G.

Para compor a supracitada base fiscal, isto é, aquela que contém apenas 1 registro por horário fiscal por dia, sugere-se trabalhar com a média aritmética simples das variáveis contínuas e a moda das categóricas em torno dos horários fiscais. Devido a possíveis incertezas quanto ao horário exato de coleta das amostras, aconselha-se considerar os

valores das variáveis preditoras 1h antes e 1h após cada horário fiscal, como exemplificado na Figura 9.

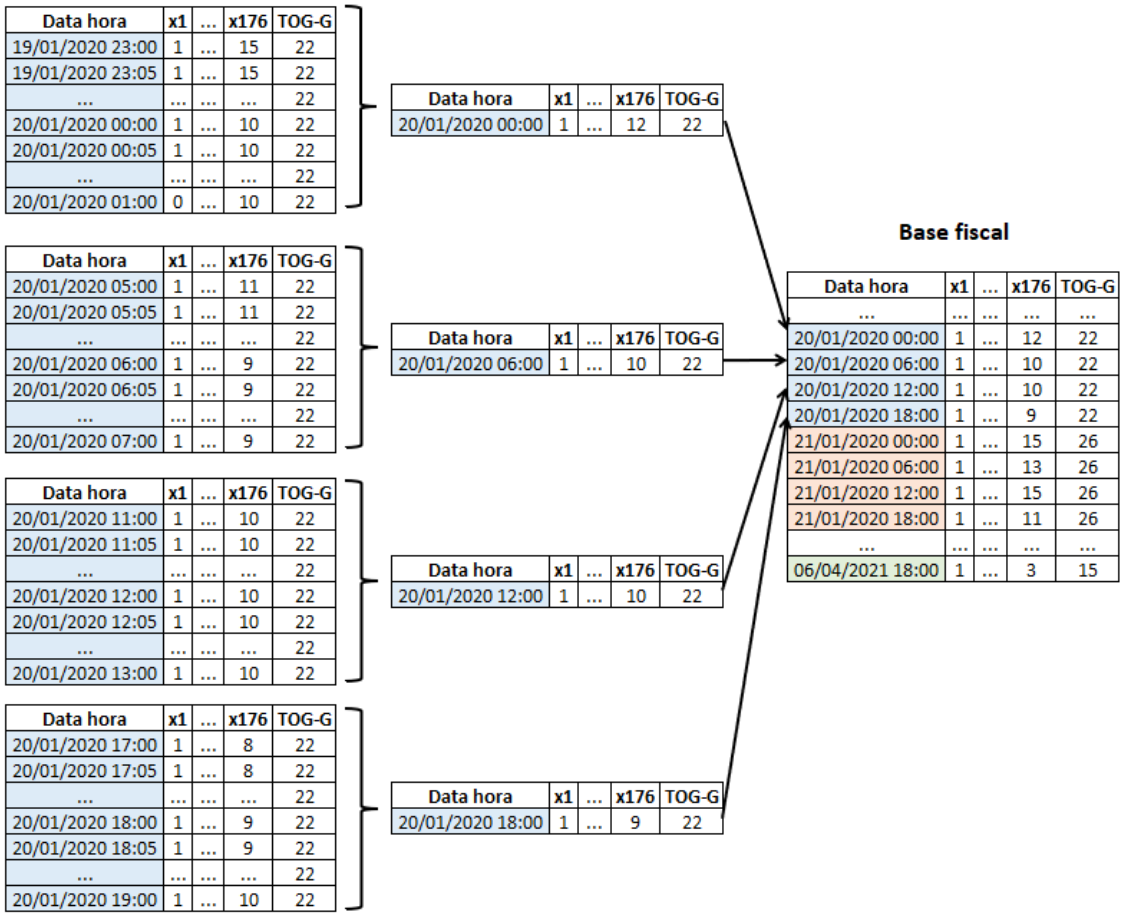


Figura 9 - Composição da base fiscal (Fonte: autoria própria)

Embora o objetivo do estudo seja estabelecer um modelo de previsão horária para o TOG-G, como as amostras coletadas nos horários fiscais são misturadas para a aferição, não temos acesso às medições do TOG-G ao longo do dia. A base fiscal não reproduz, portanto, um retrato fiel do processo de descarte, visto que, para cada registro, mantém-se o mesmo valor de TOG-G referente àquele dia. Desta forma, estas variáveis preditoras não estariam adequadas para constituir as entradas de um algoritmo de aprendizado de máquina.

Para lidar com tal obstáculo, sugere-se, então, a criação de um segundo banco de dados, denominado base diária, que deve armazenar apenas um registro por data, para a qual é possível associar um valor real de TOG-G. Para sua construção, aconselha-se a aplicação dos mesmos princípios utilizados na base fiscal: média aritmética simples das variáveis

contínuas e a moda das variáveis categóricas, porém, agora, em torno dos dias. A Figura 10 traz uma representação visual do processo de construção da base diária.

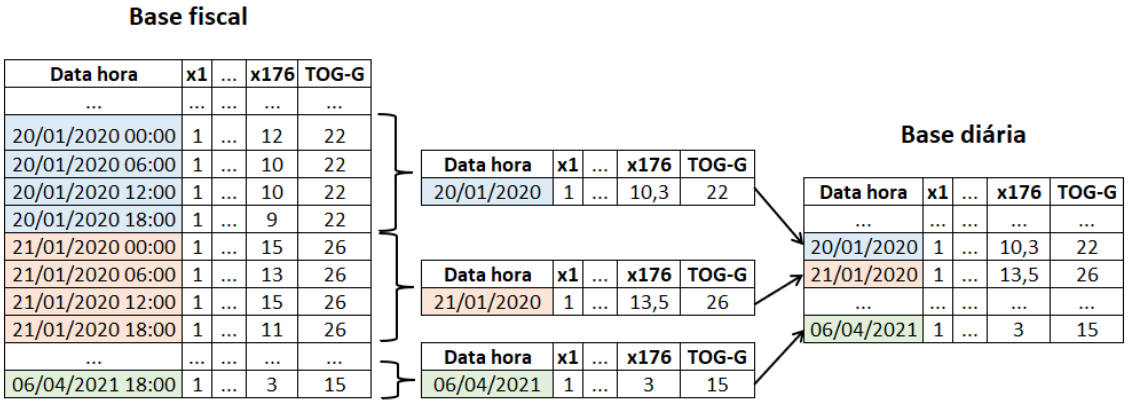


Figura 10 - Composição da base diária a partir da base fiscal (Fonte: autoria própria)

É válido destacar que, antes da extração das bases fiscal e diária, é indispensável avaliar a consistência da base histórica e dedicar um tratamento adequado às inconformidades que possam comprometer o subsequente desenvolvimento do modelo preditivo.

3.2.2. Classificação e previsão

O método proposto combina o uso de técnicas estatísticas e de aprendizado de máquina, em duas etapas, para a determinação do modelo de previsão de TOG-G. Devido à dificuldade envolvida no processo de extração de um padrão do comportamento dos dados em relação às medições de TOG-G, propõe-se neste trabalho a aplicação de uma estratégia de agrupamento de suas observações de acordo com faixas de variação, visando incorporar informações da variável de resposta ao conjunto de variáveis preditoras, com a finalidade de melhorar a acurácia da previsão do TOG-G. Desta forma, a primeira etapa do método proposto, a Etapa de Classificação - A, propõe um modelo de classificação das observações em função dos agrupamentos pré-determinados de TOG-G. Na sequência, esta informação do agrupamento juntamente das variáveis de processo e de medições de TOG aferidos por outras metodologias e disponíveis em plataforma são consideradas como entradas para o desenvolvimento do modelo de previsão do TOG-G. Para isto, assume-se que as bases de dados diária e fiscal já se encontrem pré-processadas.

3.2.2.1. Etapa de classificação - A

Passo A.1: partindo da base de dados diária, definir quantas faixas de variação do TOG-G serão extraídas. Estabelecida a quantidade de agrupamentos, definir seus limites

inferior e superior. Neste ponto, é importante considerar os valores críticos para o processo, aqui os limites de 29mg/L e 42mg/L definidos pelo CONAMA, e avaliar a sua inclusão como valores limítrofes dos agrupamentos.

Passo A.2: avaliar o balanceamento da base diária, considerando os agrupamentos definidos no Passo A.1, e aplicar a melhor técnica de balanceamento cabível, caso seja necessário. Este passo torna-se ainda mais importante quando existir desbalanceamento e o menor conjunto de observações estiver associado ao agrupamento mais crítico para o problema. Neste cenário, caso o número de registros deste agrupamento seja muito pequeno, a técnica de *oversampling* passa a ser a mais adequada. Caso contrário, ou seja, ainda que o agrupamento mais crítico seja o menor, mas ainda assim possua uma quantidade satisfatória de observações, a técnica de *undersampling* pode ser aplicada.

Passo A.3: a partir da base diária balanceada, resolver um problema de regressão linear múltipla com aplicação de algum método de *stepwise*, se necessário, considerando as variáveis preditoras e o agrupamento com o intuito de compreender e quantificar a relação entre estas entradas e a variável de resposta, identificando os preditores mais significativos para o TOG-G.

Passo A.4: reorganizar as bases diária e fiscal, mantendo apenas as variáveis com maior influência na previsão do TOG-G identificadas no Passo A.3.

Passo A.5: aplicar algoritmos de validação cruzada para dividir a base de dados diária balanceada em dois conjuntos aleatórios e distintos: treinamento e teste. É crucial que se observe a quantidade de dados da base completa para que, na extração dos conjuntos, se garanta um número mínimo viável para a execução dos testes. Uma proporção sugerida e bastante aplicada na literatura é a de 80% da base para treinamento e 20% para teste. Para a aferição da acurácia do classificador, sugere-se extrair 50 combinações diferentes de conjuntos de treinamento e teste.

Passo A.6: criar modelos de classificação do agrupamento a partir da base de dados diária, visto que, considerando a abordagem de agrupamento de TOG-G, faz-se necessário direcionar novas observações para o agrupamento adequado. Aplicar os algoritmos para cada um dos 50 pares treinamento-teste gerados no Passo A.5. Obter a acurácia de classificação a partir da média de todas as 50 acurácias de cada classificador.

Passo A.7: determinar o melhor método de classificação com base na acurácia e selecionar a combinação de conjuntos de treinamento e teste com a maior acurácia individual. Realizar a classificação das observações deste conjunto de teste eleito, associado à base diária. Avaliar o desempenho geral do classificador através do cálculo de sua acurácia, que demonstra quantas classificações corretas o modelo realizou, dentre todas as classificações geradas. Sua formulação pode ser observada na Eq. (11), onde PV = positivo verdadeiro, PF = positivo falso, NV = negativo verdadeiro e NF = negativo falso.

$$Acurácia = \frac{PV + NV}{PV + PF + NV + NF} \tag{11}$$

Passo A.8: calcular as métricas de precisão, *recall* e *f1-score* para a classificação gerada no Passo A.7.

- Precisão: verifica, dentre todas as classificações geradas pelo modelo para determinado agrupamento, quantas estão corretas. O equacionamento pode ser observado na Eq. (12).

$$Precisão = \frac{PV}{PV + PF} \tag{12}$$

- *Recall*: avalia, dentre todas as classificações esperadas para determinado agrupamento, quantas estão corretas. A Eq. (13) apresenta seu cálculo.

$$Recall = \frac{PV}{PV + NF} \tag{13}$$

- *F1-score*: estabelece a média harmônica entre precisão e *recall*, dada pela formulação apresentada na Eq. (14).

$$H = \left(\frac{\sum_{i=1}^{n} x_i^{-1}}{n} \right)^{-1} \tag{14}$$

Passo A.9: a partir das datas do conjunto de teste selecionado do Passo A.7, extrair o conjunto de teste da base fiscal e realizar a classificação de suas observações. Os agrupamentos previstos para os dados de teste das bases diária e fiscal serão entradas para a próxima etapa do método.

O fluxograma detalhado na Figura 11 sumariza os passos da Etapa de Classificação - A.

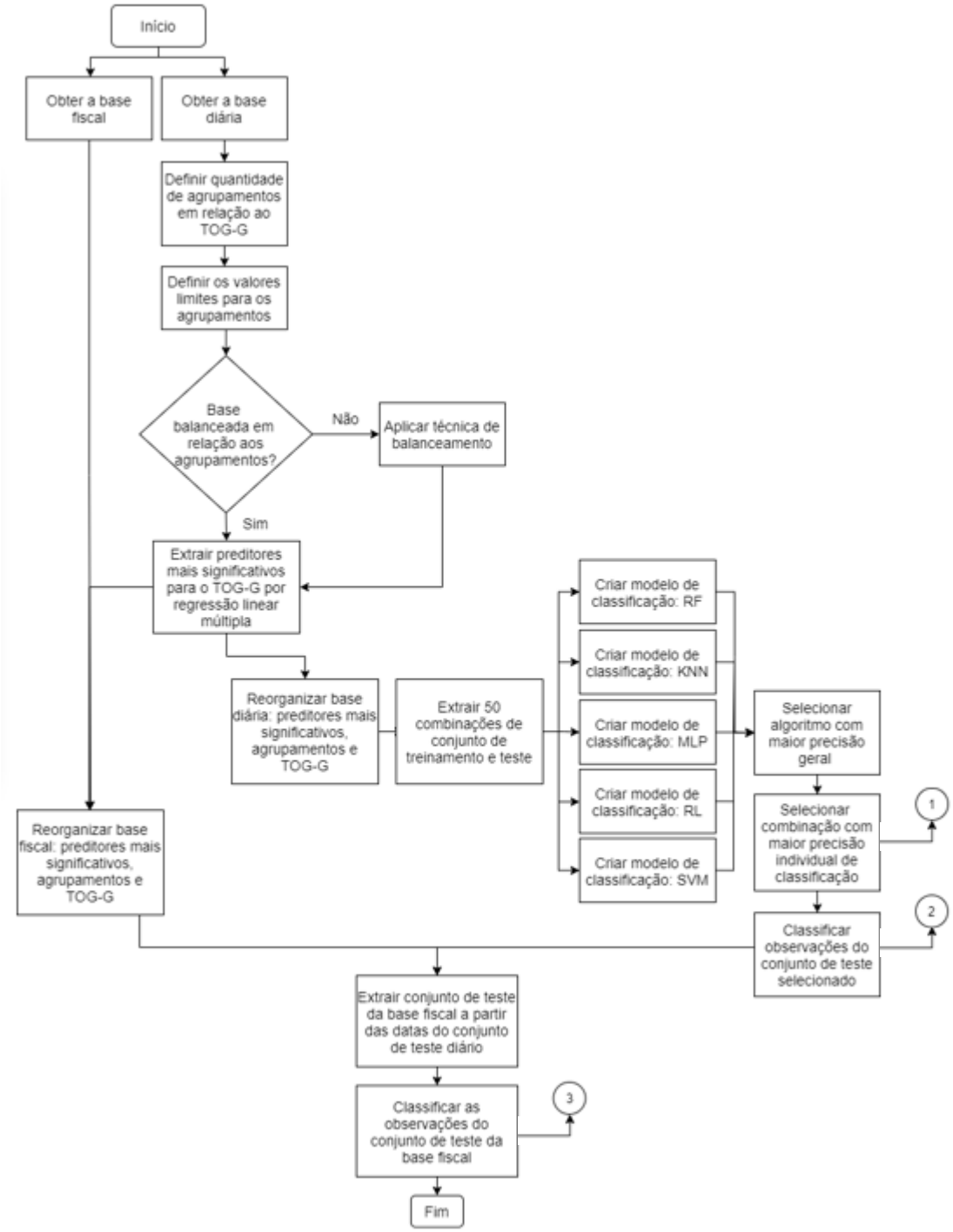


Figura 11 - Fluxograma da Etapa de Classificação – A (Fonte: autoria própria)

3.2.2.2. Etapa de previsão - B

Passo B.1: construir um modelo de previsão de TOG-G, por meio de regressão linear múltipla, para o conjunto de treinamento selecionado na etapa A.7. É importante relembrar que, neste momento, as bases já estão reorganizadas contendo apenas as variáveis preditoras significativas, conforme indicado na etapa A.4.

Passo B.2: prever valores de TOG-G para as observações de teste da base diária selecionada na etapa A.7. Para isto, considerar as previsões do conjunto de teste obtidas nesta mesma etapa A.7 como valores de entrada para a variável de agrupamento. Extrair as métricas dos valores previstos em relação aos valores reais de TOG-G:

- Erro Absoluto Médio (*Mean Absolute Error* - MAE): mensura a magnitude média dos erros, em módulo, do conjunto de previsões, de acordo com a Eq. (15) onde y_i são os valores previstos, enquanto x_i constituem os valores reais e n representa a quantidade de observações.

$$MAE = \frac{\sum_{i=1}^{n} |y_i - x_i|}{n} \tag{15}$$

- Erro Percentual Absoluto Médio (*Mean Absolute Percentage Error* - MAPE): expressa o percentual médio dos erros, em módulo, do conjunto de previsões. Sua formulação pode ser observada na Eq. (16), onde x_i e y_i representam os valores reais e os valores previstos, respectivamente, e n a quantidade de observações.

$$MAPE = \frac{100\%}{n} \sum_{i=1}^{n} \left| \frac{x_i - y_i}{x_i} \right| \tag{16}$$

- Coeficiente de determinação (R^2): representa o ajuste de um modelo estatístico linear generalizado em relação aos valores reais de uma variável aleatória. Seu cálculo pode ser observado na Eq. (17), onde n refere-se à quantidade de observações, x_i representa os valores reais, y_i os valores previstos e $\bar{x}$ a média dos valores reais.

$$R^2 = 1 - \frac{\sum_{i=1}^{n} (x_i - y_i)^2}{\sum_{i=1}^{n} (x_i - \bar{x})^2} \tag{17}$$

- Coeficiente de correlação de Pearson (ρ): mede o grau e a direção da relação linear entre duas variáveis. Seu valor pode variar de -1 a 1, e segue a formulação apresentada na Eq. (18), onde n refere-se à quantidade de observações, x_i representa os valores reais, y_i os valores previstos, $\bar{x}$ a média dos valores reais e $\bar{y}$ a média dos valores previstos.

$$\rho_{x,y} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})((y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2 \sum_{i=1}^{n}(y_i - \bar{y})^2}} \quad (18)$$

Passo B.3: prever os valores de TOG-G para o conjunto de teste da base fiscal extraído na etapa A.9. Considerar as classificações para o conjunto de teste da base fiscal obtidas na etapa A.9. como entradas para o modelo.

Passo B.4: extrair a média e o terceiro quartil, por dia, dos valores de TOG-G previstos na etapa B.3. Como não existem medidas reais de TOG-G para cada horário fiscal, não é possível realizar o confronto direto com os resultados previstos. Entretanto, esta previsão horária é a de maior interesse neste estudo, conforme previamente mencionado. Assim, como o valor de TOG-G é medido a partir de uma amostra que mistura as 4 porções coletadas nos 4 horários fiscais, julgou-se adequado extrair a média e o terceiro quartil dos dados previstos por dia, como medidas de comparação com os valores de TOG-G reais do conjunto de teste diário. Extrair as métricas MAE, MAPE, R^2 e ρ dos valores de média e terceiro quartil dos valores previstos em relação aos valores reais de TOG-G.

O fluxograma da Figura 12 apresenta os passos pertencentes à Etapa de Previsão - B.

3.2.3. Confirmação

Considerando a inovação trazida pelo método, de incorporação de informações da variável a ser prevista ao conjunto de preditoras utilizando um modelo de classificação gerado pela aplicação de técnicas de *machine learning*, esta etapa do método visa confirmar as melhorias nos resultados obtidos quando confrontados com os resultados produzidos por modelos de regressão linear múltipla construídos sem a presença da informação do agrupamento. Para tal, sugere-se a criação de um novo modelo a partir das variáveis preditoras mais significativas selecionadas na etapa A.3 e o posterior cálculo das mesmas métricas aplicadas durante a etapa de Previsão.

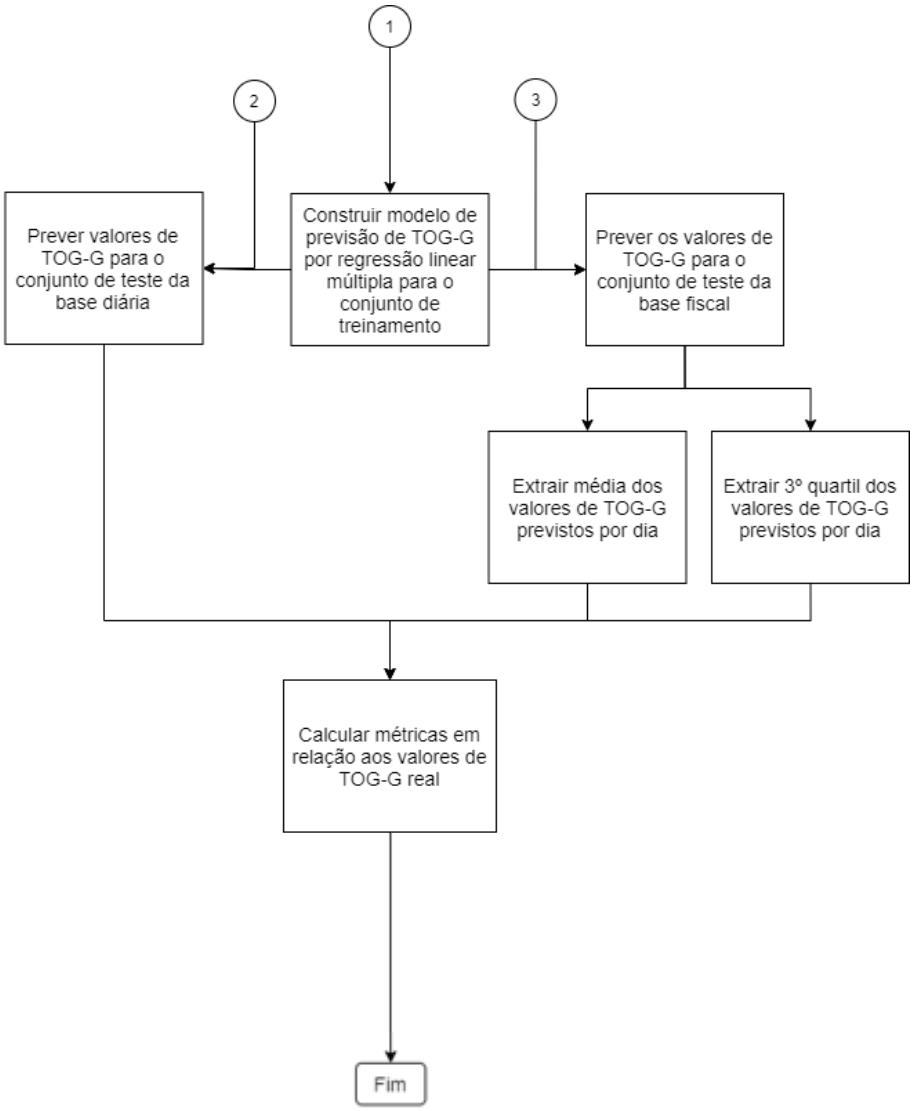


Figura 12 - Fluxograma da Etapa de Previsão – B (Fonte: autoria própria)

Além disso, como a plataforma fornece uma medida de referência alternativa ao TOG-G para uso em tempo real, o TOG espectrofotométrico, esta informação também deve ser considerada para avaliação do método. Portanto, igualmente sugere-se o cálculo das métricas de TOG-S *versus* TOG-G. Desta forma, torna-se possível quantificar os ganhos obtidos com o método proposto.

3.3. Caracterização da pesquisa

Considerando as diretrizes de classificação de pesquisas científicas em Engenharia de Produção (MIGUEL *et al.*, 2014), este estudo caracteriza-se como de natureza aplicada devido ao seu interesse prático e aplicado a problemas reais, e de abordagem quantitativa,

visto que investiga a influência de fatores na ocorrência de determinado fenômeno a partir de dados que podem ser quantificados e analisados estatisticamente (APPOLINARIO, 2009; GERHARDT; SILVEIRA, 2019). Neste sentido, traz como método de pesquisa a modelagem, por meio da qual a complexidade do problema real é traduzida em modelos matemáticos; com característica normativa em termos de objetivo, por buscar uma solução em um delineamento definido a partir de técnicas disponíveis na literatura (BERTRAND; FRANSOO, 2002).

4. APLICAÇÃO DO MÉTODO

4.1. Conjunto de dados de Teor de Óleos e Graxas (TOG)

4.1.1. Considerações iniciais

Os dados considerados neste estudo foram obtidos de uma plataforma FPSO para processamento primário de petróleo. Nesse tipo de processamento, uma grande parte da água produzida que é gerada pode ser despejada no mar desde que atenda a determinados requisitos legais. Conforme mencionado na Seção 2, atualmente a agência brasileira regulamentadora prevê que o valor de TOG-G diário deve ser inferior a 42 mg/L e sua média mensal inferior a 29 mg/L.

4.1.2. Pré-processamento

O conjunto de dados original possuía 314.208 registros referentes ao histórico de um período não contínuo aproximado de 2 anos, com informações relacionadas a 174 variáveis de processo, além das medidas de TOG *eracheck* (TOG-E), TOG espectrofotométrico e TOG gravimétrico. Dentre as variáveis, 13 eram categóricas e 164 contínuas. O Apêndice A apresenta a descrição das 177 variáveis bem como indica se a variável é controlável ou não, dentro do processo. Com exceção das medidas de TOG-G que são diárias, os demais dados foram coletados em intervalos de 5 em 5 minutos.

Para a estruturação da base fiscal, observaram-se os horários fiscais determinados para esta plataforma: 00h, 06h, 12h e 18h. O Quadro 1 apresenta a composição dos registros da base fiscal, considerando os valores das variáveis preditoras 1h antes e 1h após cada horário fiscal. É válido ressaltar que, sendo o primeiro horário à meia noite, para a composição do primeiro registro fiscal de D-0 da nova base, os valores referentes às últimas 12 medições registradas no dia anterior (D-1) foram considerados.

Quadro 1 - Composição dos registros da base fiscal para o dia D-0

	D-1	**D-0**
00h	23:00 – 23:55	00:00 – 01:00
06h	-	05:00 – 07:00
12h	-	11:00 – 13:00
18h	-	17:00 – 19:00

Calcularam-se a média aritmética simples das variáveis contínuas e a moda das categóricas para os intervalos de dados relacionados aos horários fiscais. A estruturação final da base fiscal apresentou 4.196 registros referentes a 1.049 datas.

Na sequência e da mesma forma, calculando-se a média aritmética simples das variáveis contínuas e a moda das categóricas, agora para cada data da base fiscal, obteve-se a base diária com 1.049 registros em sua estrutura, conforme esperado.

Antes da extração das bases fiscal e diária a partir do banco de dados original, porém, foram descartadas as linhas que não continham informação de TOG-G, visto que este constitui dado essencial para a análise proposta. Dados inconsistentes apresentando valores negativos foram substituídos por valores faltantes, para não comprometer o cálculo das médias. Além disso, as variáveis categóricas (x_1 até x_{13}), relacionadas a métodos de elevação, foram codificadas segundo o Quadro 2.

Quadro 2 - Codificação das variáveis categóricas

Valor original	Valor codificado
BCS	1
BCSS	1
GLC	1
Gas Lift	1
Surgente	1
Fechado	0

A compilação de ambas bases foi realizada por meio de programação computacional utilizando a linguagem Python.

4.1.3. Pré-análise

De posse da base fiscal e da base diária, realizou-se uma nova análise de valores faltantes, dado que sua presença interfere na criação de um modelo preditivo adequado, alvo deste estudo. Assim, detectou-se a presença de valores faltantes em 33 variáveis do conjunto diário, cujos percentuais estão destacados na Tabela 1.

Tabela 1 - Quantidade de valores faltantes na base diária

Código da variável	Quantidade de valores faltantes	Percentual de valores faltantes
x_{50}	21	2,00%
x_{51}	119	11,34%
x_{54}	228	21,73%
x_{56}	60	5,72%
x_{58}	120	11,44%
x_{59}	1	0,10%
x_{61}	432	41,18%
x_{62}	817	77,88%
x_{63}	691	65,87%
x_{64}	991	94,47%
x_{65}	97	9,25%
x_{66}	814	77,60%
x_{68}	464	44,23%
x_{69}	899	85,70%
x_{73}	45	4,29%
x_{77}	59	5,62%
x_{79}	6	0,57%
x_{80}	13	1,24%
x_{81}	277	26,41%
x_{82}	582	55,48%
x_{83}	901	85,89%
x_{84}	421	40,13%
x_{85}	628	59,87%
x_{102}	91	8,67%
x_{103}	33	3,15%
x_{104}	2	0,19%
x_{109}	60	5,72%
x_{110}	35	3,34%
x_{111}	4	0,38%
x_{113}	24	2,29%
x_{115}	200	19,07%
x_{117}	255	24,31%
x_{175}	679	65%

Na sequência, avaliaram-se as correlações existentes entre as variáveis de processo e as variáveis TOG-E e TOG-S com a variável cujo valor deseja-se prever, TOG-G. O Apêndice B apresenta os valores de correlação e seus intervalos de confiança associados,

bem como seus respectivos *p-values*. Foi possível observar que mais de 90% das variáveis preditoras possui correlação baixa ($\rho < |0{,}5|$) e não significativa ($p\text{-}value < 0{,}05$) com a variável de resposta.

Considerando a complexidade de análise envolvida e a qualidade das bases de dados disponíveis, decidiu-se por ouvir a opinião de especialistas na área de Petróleo, no sentido de pré-selecionar as variáveis de entrada dentro de seu impacto para o processo. Das 176 variáveis originais, 80 foram consideradas relevantes, além do TOG-E e TOG-S, e estão apresentadas no Quadro 3.

Quadro 3 - Variáveis selecionadas a partir da opinião especializada

Código	Descrição	Código	Descrição
x_{49}	Corrente Trafo B do Pré-TO	x_{121}	Total Água Produzida Poço 3
x_{50}	Corrente Trafo C do Pré-TO	x_{122}	Total Água Produzida Poço 4
x_{51}	Corrente Trafo A TO	x_{123}	Total Água Produzida Poço 5
x_{52}	Corrente Trafo B TO	x_{124}	Total Água Produzida Poço 6
x_{53}	Corrente Trafo C TO	x_{125}	Total Água Produzida Poço 7
x_{54}	Nível óleo *settling* B	x_{126}	Total Água Produzida Poço 8
x_{55}	Nível separador de produção HP	x_{127}	Total Água Produzida Poço 9
x_{56}	Nível interface Pré-TO	x_{128}	Total Água Produzida Poço 10
x_{57}	Nível separador de produção LP	x_{129}	Total Água Produzida Poço 11
x_{58}	Nível interface TO	x_{130}	Total Água Produzida Poço 12
x_{73}	Pressão separador de produção LP	x_{131}	Total Água Produzida Poço 13
x_{74}	Pressão TO	x_{132}	BSW Poço 1
x_{76}	Pressão separador de produção HP	x_{133}	BSW Poço 2
x_{77}	Pressão Pré-TO	x_{134}	BSW Poço 3
x_{86}	Pressão saída de óleo TO	x_{135}	BSW Poço 4
x_{89}	Temperatura entrada separador de produção HP	x_{136}	BSW Poço 5

(continuação)

Código	Descrição	Código	Descrição
x_{90}	Temperatura TO	x_{137}	BSW Poço 6
x_{94}	Temperatura entrada separador de produção LP	x_{138}	BSW Poço 7
x_{96}	Temperatura do *settling* tanque 4 central	x_{139}	BSW Poço 18
x_{97}	Temperatura do *settling* tanque 3 central	x_{140}	BSW Poço 9
x_{99}	TOG no descarte da água produzida	x_{141}	BSW Poço 10
x_{101}	Vazão de água produzida para hidrociclone A	x_{142}	BSW Poço 11
x_{102}	Vazão de água produzida entrada hidrociclone B	x_{143}	BSW Poço 12
x_{103}	Nível na câmara de óleo do flotador A	x_{144}	BSW Poço 13
x_{104}	Nível na câmara de óleo do flotador B	x_{158}	Produção Líquida de Óleo Poço 1
x_{105}	Nível da câmara de água flotador A	x_{159}	Produção Líquida de Óleo Poço 2
x_{106}	Nível da câmara de água flotador B	x_{160}	Produção Líquida de Óleo Poço 3
x_{107}	Tempo de residência flotador A	x_{161}	Produção Líquida de Óleo Poço 4
x_{108}	Tempo de residência flotador B	x_{162}	Produção Líquida de Óleo Poço 5
x_{109}	Pressão diferencial entrada/saída de água hidrociclone A	x_{163}	Produção Líquida de Óleo Poço 6
x_{110}	Pressão diferencial entrada/saída de água hidrociclone B	x_{164}	Produção Líquida de Óleo Poço 7
x_{111}	Pressão flotador A	x_{165}	Produção Líquida de Óleo Poço 8
x_{112}	Pressão flotador B	x_{166}	Produção Líquida de Óleo Poço 9
x_{113}	Pressão flotador A	x_{167}	Produção Líquida de Óleo Poço 10
x_{114}	Pressão flotador B	x_{168}	Produção Líquida de Óleo Poço 11
x_{115}	Desemulsificante *Topside*	x_{169}	Produção Líquida de Óleo Poço 12
x_{116}	Desemulsificante *Subsea*	x_{170}	Produção Líquida de Óleo Poço 13
x_{117}	Inibidor de incrustação *Topside*	x_{172}	Total Água Produzida
x_{118}	Polieletrólito	x_{174}	Total Produção Líquida de Óleo

(conclusão)

Código	Descrição	Código	Descrição
x_{119}	Total Água Produzida Poço 1	x_{175}	TOG *eracheck*
x_{120}	Total Água Produzida Poço 2	x_{176}	TOG espectrofotométrico

Embora tenha sido indicada pelos especialistas, a variável x_{175} referente ao TOG-E foi desconsiderada para análise, pois avaliou-se que a disponibilidade de seus dados históricos estava com período incompatível para com os demais dados. Trabalhar com esta informação na base, faria com que o conjunto fosse reduzido aos últimos 370 registros, dentre os quais apenas 3 observações conteriam informação relevante de desenquadramento de TOG-G.

Para a aplicação do método experimental, portanto, foram consideradas as bases de dados diária e fiscal compostas pelas 80 variáveis de entrada elencadas nesta subseção, além do TOG-S.

4.1.4. Etapa de Classificação - A

Passo A.1: para definição dos agrupamentos, plotou-se um *boxplot* para os dados de TOG-G da base diária, a partir do qual obtiveram-se os valores 10, 14 e 20 para o primeiro quartil, mediana e terceiro quartil, respectivamente, conforme as linhas pontilhadas em vermelho da Figura 13.

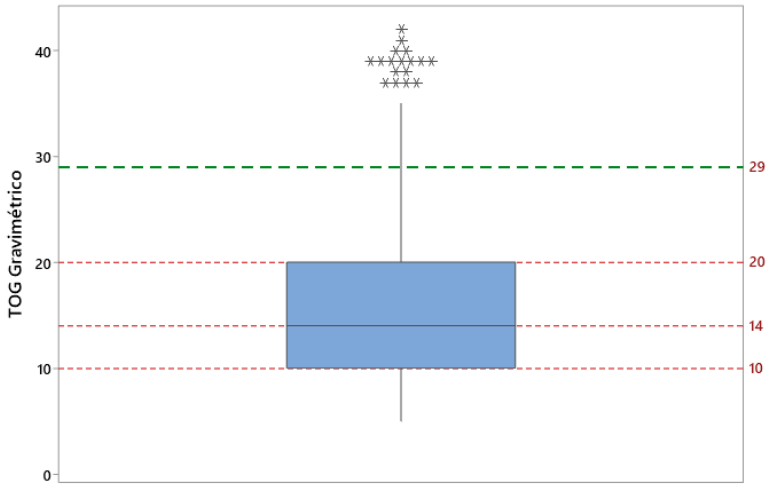


Figura 13 - *Boxplot* do TOG-G (Fonte: autoria própria)

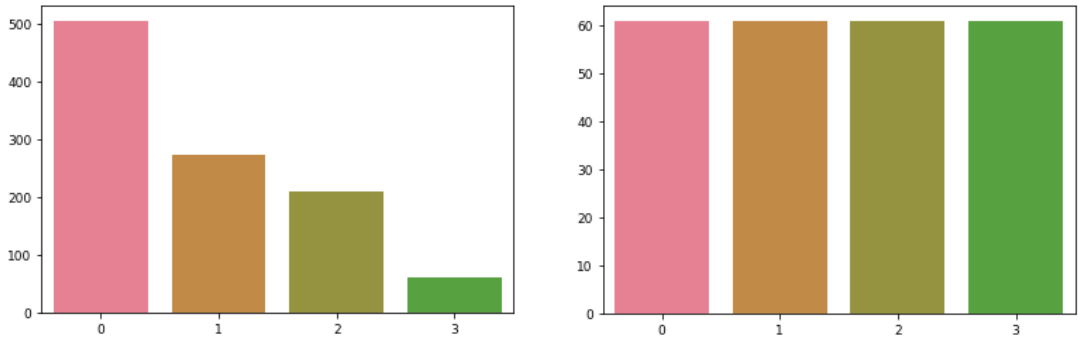
No entanto, conforme mencionado no Capítulo 2, os valores limites diário e mensal para o TOG-G são 42mg/L e 29mg/L, respectivamente. Optou-se, então, por não considerar o primeiro quartil como limite para o primeiro agrupamento, visto que valores inferiores a 10 não eram o de maior interesse ao contexto de análise. Por outro lado, e mantendo uma abordagem conservadora, valores acima do limite médio mensal indicado pelo CONAMA, conforme indicado na linha tracejada verde da Figura 13, seriam os de maior interesse dentro do conjunto. Assim, estabeleceram-se os limites inferior e superior para cada agrupamento, e o total de registros resultantes para cada um deles pode ser observado na Tabela 2.

Tabela 2 - Limites inferior e superior e quantidade de registros para cada agrupamento

Agrupamento	Limite inferior	Limite superior	Quantidade de registros
0	0	13	505
1	14	19	273
2	20	28	210
3	29	42	61

Passo A.2: uma vez que os valores dos quartis não foram totalmente considerados para o agrupamento das observações, produziu-se um conjunto de dados desequilibrado em relação à quantidade de registros em cada agrupamento. É possível observar na Tabela 2 e na Figura 14(a) que o agrupamento 3, que contempla as informações de maior interesse para previsão dentro deste estudo – os picos de TOG-G –, possui a menor quantidade de dados. Desta forma, aplicou-se a técnica de subamostragem (*undersampling*) a fim de

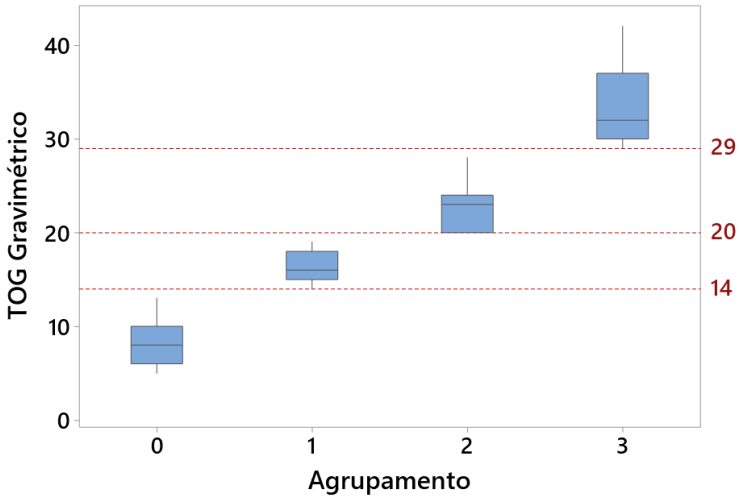
balancear a base, resultando na composição igualitária de 61 registros para cada agrupamento, conforme mostrado nas Figura 14(b), Figura 15(a) e Figura 15(b).



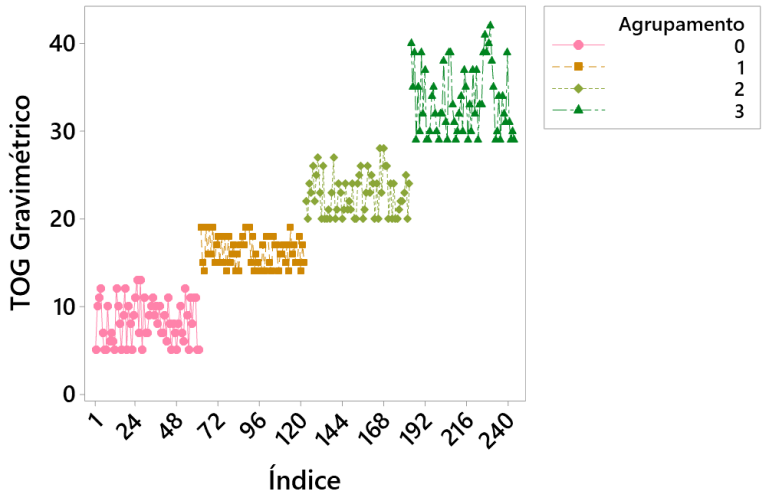
(a) (b)

Figura 14 - Conjunto de dados desbalanceado (a) e balanceado (b) através do *undersampling* (Fonte: autoria própria)

Essa abordagem garantiu que as informações reais do agrupamento 3 fossem mantidas, enquanto que os melhores dados para os demais fossem selecionados.



(a)



(b)

Figura 15 - *Boxplot* (a) e *time series plot* (b) do TOG-G para o conjunto de dados balanceado (Fonte: autoria própria)

Passo A.3: a partir da base diária balanceada, que totalizou 244 registros em sua composição, resolveu-se um problema de regressão linear múltipla, considerando as 80 variáveis originais indicadas pelos especialistas, o TOG-S e o agrupamento. Foi aplicado um *Backward Elimination* com $\alpha = 0{,}05$ a fim de reduzir o número de variáveis no modelo e aumentar sua capacidade preditiva. As variáveis selecionadas na regressão foram mantidas durante toda a solução do problema. A Tabela 3 apresenta a análise de variância (ANOVA) relativa a esta análise.

Tabela 3 - ANOVA das variáveis mais significativa para a previsão do TOG-G

Fonte de variação	GL	SQ	SQA	QMA	*F-value*	*p-value*
Regressão	11	21120,8	21120,8	1920,07	329,99	0,000
x_{101}	1	313,0	43,6	43,62	7,50	0,007
x_{108}	1	109,6	52,9	52,85	9,08	0,003
x_{122}	1	93,6	84,6	84,56	14,53	0,000
x_{128}	1	7707,5	108,7	108,67	18,68	0,000
x_{161}	1	82,2	129,7	129,67	22,29	0,000
x_{162}	1	5,1	30,00	30,00	5,16	0,024
x_{168}	1	240,1	122,9	122,92	21,13	0,000
x_{176}	1	1710,0	68,4	68,40	11,76	0,001
Agrupamento	3	10859,8	10859,8	3619,93	622,14	0,000
Erro	183	1349,9	1349,9	5,82		
Total	194	22470,7				

As variáveis selecionadas para a previsão do TOG-G, portanto, foram: vazão de água produzida para hidrociclone A, tempo de residência no flotador B, total de água produzida Poço 4, total de água produzida Poço 10, produção líquida de óleo Poço 4, produção líquida de óleo Poço 5, produção líquida de óleo Poço 11 e TOG-S. É possível observar na Tabela 4 a correlação de Pearson entre estas variáveis e o TOG-G para as observações da base diária balanceada, visualmente representada por escala de cores no correlograma plotado na Figura 16.

Tabela 4 - Correlação de Pearson entre as variáveis preditoras selecionadas e o TOG-G

Variável	Correlação	95% IC para ρ	*p-value*
x_{101}	0,118	(-0,008; 0,240)	0,066
x_{108}	0,047	(-0,079; 0,172)	0,464
x_{122}	0,069	(-0,057; 0,193)	0,286
x_{128}	-0,561	(-0,642; -0,469)	0,000
x_{161}	0,277	(0,156; 0,389)	0,000
x_{162}	-0,126	(-0,247; 0,000)	0,050
x_{168}	0,199	(0,075; 0,316)	0,002
x_{176}	0,554	(0,461; 0,635)	0,000

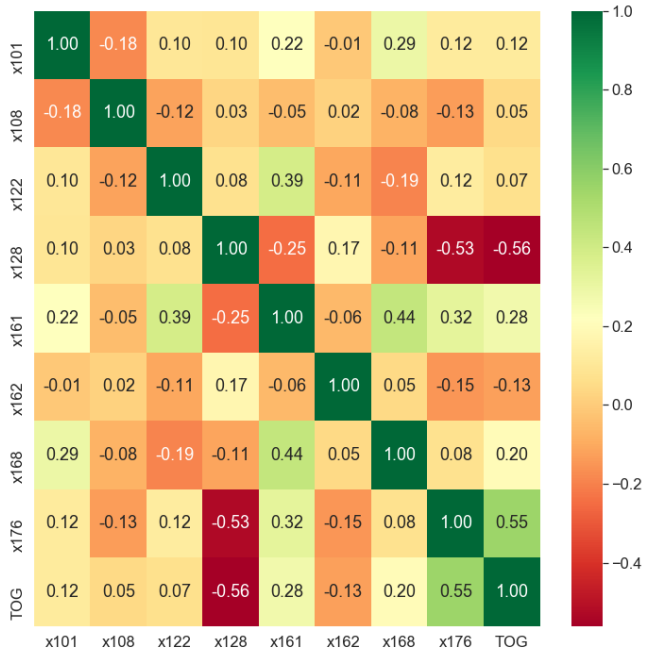


Figura 16 - Correlograma para as variáveis preditoras selecionadas e TOG-G (Fonte: autoria própria)

Passo A.4: a partir da análise realizada na etapa anterior, as bases diária e fiscal foram reorganizadas para que fossem mantidas apenas as variáveis preditoras mais significativas.

Passo A.5: a base diária balanceada foi dividida em 50 pares distintos de conjuntos de dados de treinamento e teste, seguindo a proporção de 80/20, respectivamente. A seleção de observações foi realizada por meio de programação computacional utilizando a função de validação cruzada *Stratified K-Folds* da linguagem Python.

Passo A.6: Para a construção do classificador de agrupamentos, aplicaram-se os algoritmos *Random Forest Classifier*, *K-Nearest Neighbors*, *Multilayer Perceptron*, Regressão Logística e *Support Vector Machine*, através de programação computacional utilizando a linguagem Python. Definiu-se como variáveis de entrada o mesmo conjunto determinado no Passo A.3.

A Tabela 5 apresenta os parâmetros utilizados para os cinco algoritmos que foram aplicados aos 50 conjuntos de treinamento extraídos, bem como suas acurácias resultantes. Para que fosse possível reproduzir os resultados obtidos, manteve-se o $random\ state = 0$ durante a execução dos programas.

Tabela 5 - Parâmetros e acurácias dos algoritmos de classificação

Algoritmo	Parâmetros	Acurácia
RF	$n_{estimators} = 150$ $max_{features} = \sqrt{n_{features}} = 3$	48,61%
KNN	$n_{neighbors} = 5$ $p = 2$	38,67%
MLP	$hiddenlayer_{sizes} = (50,50,50)$ $activation = 'relu'$ $solver = 'adam'$ $alpha = 0{,}0001$ $learning_{rate} = 'invscaling'$	36,05%
RL	$max_{iter} = 1000$	45,56%
SVM	$kernel = 'rbf'$ $C = 2$	44,40%

Passo A.7: com a melhor acurácia obtida, 48,61%, determinou-se o algoritmo *Random Forest* para compor o classificador a ser utilizado ao longo da solução desta subseção. Selecionou-se, então, a combinação de conjuntos de treinamento e teste com a maior acurácia individual para o algoritmo RF, de maneira que fosse possível realizar a classificação das observações dos conjuntos de teste associados à base diária e fiscal.

A composição do par treinamento-teste selecionado em relação aos 4 agrupamentos de TOG-G estabelecidos pode ser observada na Tabela 6.

Tabela 6 - Distribuição dos agrupamentos dentro dos conjuntos de treinamento e teste

Agrupamento	Conjunto de Treinamento	Conjunto de teste
0	50	17
1	48	11
2	51	10
3	46	11

Aplicando o classificador ao conjunto de teste, associado à base diária, obtiveram-se os valores apresentados na Tabela 7. Para facilitar a associação posterior do conjunto de teste da base diária com o conjunto de teste da base fiscal, as datas foram codificadas em numeração sequencial.

Tabela 7 - Classificações de agrupamento para o conjunto de teste da base diária

Data	Agrupamento real	Agrupamento previsto
1	0	0
2	3	3
3	0	0
4	3	3
5	1	0
6	1	0
7	1	1
8	2	2
9	3	3
10	0	0
11	2	0
12	0	1
13	3	3
14	0	0
15	3	1
16	1	1
17	2	2
18	1	2
19	0	0
20	3	1

(conclusão)

Data	Agrupamento real	Agrupamento previsto
21	0	0
22	0	0
23	0	0
24	1	1
25	3	2
26	3	3
27	1	1
28	0	0
29	3	3
30	1	1
31	0	0
32	2	1
33	3	3
34	3	1
35	2	2
36	1	2
37	2	1
38	1	2
39	3	0
40	2	3
41	1	0
42	2	3
43	3	3
44	1	0
45	3	2
46	2	2
47	1	0
48	3	3
49	2	2

Passo A.8: a matriz de confusão e as métricas de precisão, *recall* e *f1-score* relacionadas a este resultado estão apresentadas na Tabela 8 e Tabela 9, respectivamente.

Tabela 8 - Métricas de avaliação para as classificações da base diária

Agrupamento	Precisão	*Recall*	*f-1 score*
0	0,62	0,91	0,74
1	0,45	0,38	0,42
2	0,50	0,60	0,55
3	0,90	0,60	0,72
Acurácia			0,61

Dado que esses valores previstos de agrupamento serão posteriormente utilizados como variável categórica para prever as medições de TOG-G, é importante destacar que os erros de classificação dos agrupamentos do triângulo superior não são a principal preocupação,

uma vez que um TOG-G previsto maior do que realmente é não representa um sério inconveniente. Por outro lado, a previsão de um valor baixo para um TOG-G real fora das especificações não alertaria para a necessidade de medidas preventivas ou corretivas, podendo levar a prejuízos importantes.

Tabela 9 - Matriz de confusão para a previsão do Agrupamento do conjunto de testes da base diária

		Agrupamento Previsto				
		0	**1**	**2**	**3**	**Total**
Agrupamento Real	**0**	10	1	0	0	11
	1	5	5	3	0	13
	2	0	3	6	1	10
	3	1	2	3	9	15
	Total	16	11	12	10	49

Passo A.9: a partir das datas selecionadas para o conjunto de teste da base diária indicada no passo imediatamente anterior, extraiu-se o conjunto de teste da base fiscal, de maneira que para cada registro da base diária estivessem associadas 4 observações da base fiscal (quatro horários fiscais para a mesma data). Aplicando o classificador ao conjunto de teste, agora relacionado à base fiscal, obtiveram-se os agrupamentos apresentados na Tabela 10. A sequência de observações associada a cada data apresentada na Tabela 10 será mantida ao longo de todo o trabalho.

Tabela 10 - Classificações de agrupamento e probabilidade associada a cada uma das classes o conjunto de teste da base fiscal

Data	**Agrupamento previsto**	**Probabilidade de a observação pertencer à classe:**			
		0	1	2	3
1	0	62%	33%	3%	1%
1	0	63%	28%	3%	6%
1	0	57%	35%	5%	2%
1	0	58%	35%	5%	2%
2	**3**	1%	19%	21%	59%
2	**2**	1%	15%	53%	31%
2	**2**	1%	15%	50%	35%
2	**3**	1%	14%	33%	52%
3	0	57%	42%	0%	1%
3	0	53%	46%	0%	1%
3	0	55%	44%	0%	1%
3	0	55%	44%	0%	1%
4	3	2%	23%	19%	56%
4	3	3%	35%	18%	44%
4	3	7%	31%	19%	43%

(continuação)

Data	Agrupamento previsto	Probabilidade de a observação pertencer à classe:			
		0	1	2	3
4	3	7%	33%	19%	42%
5	1	12%	33%	29%	27%
5	1	39%	43%	9%	9%
5	1	41%	41%	9%	9%
5	1	40%	41%	9%	9%
6	**1**	46%	53%	0%	1%
6	**0**	62%	37%	0%	1%
6	**0**	62%	37%	0%	1%
6	**0**	62%	37%	0%	1%
7	**2**	16%	30%	37%	17%
7	**1**	12%	37%	33%	18%
7	**2**	15%	23%	34%	28%
7	**3**	3%	21%	37%	39%
8	2	2%	11%	63%	24%
8	2	3%	17%	58%	22%
8	2	5%	21%	52%	22%
8	2	3%	12%	63%	22%
9	3	1%	11%	31%	57%
9	3	3%	7%	30%	60%
9	3	3%	4%	33%	61%
9	3	5%	11%	32%	53%
10	0	87%	12%	0%	1%
10	0	84%	12%	0%	4%
10	0	84%	12%	0%	4%
10	0	84%	12%	0%	4%
11	1	19%	35%	24%	22%
11	1	13%	56%	23%	8%
11	1	13%	59%	21%	7%
11	1	4%	43%	27%	26%
12	**1**	8%	40%	28%	24%
12	**1**	4%	59%	27%	11%
12	**2**	2%	31%	41%	26%
12	**2**	1%	16%	62%	21%
13	3	3%	23%	19%	55%
13	3	3%	21%	29%	48%
13	3	3%	21%	20%	55%
13	3	3%	23%	23%	51%
14	0	85%	14%	0%	1%
14	0	86%	13%	0%	1%
14	0	76%	23%	0%	1%
14	0	77%	21%	0%	1%
15	1	2%	65%	21%	12%
15	1	9%	45%	26%	21%
15	1	2%	44%	37%	17%
15	1	2%	45%	31%	22%
16	1	15%	66%	2%	17%
16	1	16%	65%	3%	15%
16	1	18%	63%	4%	15%
16	1	17%	64%	3%	16%
17	2	3%	19%	43%	35%

(continuação)

Data	Agrupamento previsto	Probabilidade de a observação pertencer à classe:			
		0	1	2	3
17	2	3%	26%	45%	27%
17	2	0%	13%	56%	31%
17	2	2%	21%	51%	26%
18	2	19%	16%	58%	7%
18	2	19%	15%	60%	7%
18	2	20%	17%	57%	6%
18	2	21%	16%	57%	6%
19	0	51%	17%	22%	11%
19	0	61%	13%	16%	10%
19	0	66%	14%	12%	8%
19	0	51%	14%	21%	14%
20	1	3%	47%	29%	21%
20	1	3%	54%	22%	21%
20	1	2%	47%	33%	19%
20	1	1%	59%	16%	23%
21	0	58%	29%	9%	4%
21	0	48%	21%	21%	10%
21	0	47%	17%	22%	14%
21	0	47%	17%	22%	14%
22	0	63%	19%	12%	5%
22	0	50%	24%	20%	6%
22	0	53%	21%	21%	5%
22	0	51%	23%	21%	5%
23	0	81%	16%	1%	1%
23	0	57%	39%	1%	2%
23	0	57%	39%	1%	2%
23	0	57%	39%	1%	2%
24	1	13%	77%	1%	9%
24	1	13%	77%	1%	9%
24	1	14%	76%	1%	9%
24	1	14%	76%	1%	9%
25	**2**	9%	13%	39%	39%
25	**3**	20%	12%	27%	41%
25	**3**	5%	15%	39%	40%
25	**2**	1%	19%	55%	25%
26	3	1%	17%	15%	66%
26	3	1%	15%	20%	63%
26	3	1%	14%	23%	63%
26	3	1%	17%	17%	64%
27	**0**	53%	46%	0%	1%
27	**1**	48%	50%	1%	1%
27	**1**	47%	51%	1%	1%
27	**1**	48%	50%	1%	1%
28	0	72%	13%	5%	11%
28	0	64%	19%	10%	7%
28	0	67%	16%	9%	8%
28	0	72%	15%	5%	8%
29	**3**	3%	16%	22%	59%
29	**1**	4%	46%	15%	35%

(continuação)

Data	Agrupamento previsto	Probabilidade de a observação pertencer à classe:			
		0	1	2	3
29	**3**	7%	18%	14%	61%
29	**3**	10%	29%	15%	46%
30	1	0%	59%	21%	21%
30	1	1%	50%	19%	31%
30	1	1%	47%	25%	28%
30	1	0%	53%	26%	21%
31	**1**	41%	59%	0%	1%
31	**0**	54%	45%	1%	1%
31	**0**	54%	45%	1%	1%
31	**0**	63%	27%	0%	9%
32	**3**	3%	23%	20%	54%
32	**2**	1%	33%	41%	25%
32	**1**	1%	39%	32%	28%
32	**1**	2%	36%	34%	28%
33	3	1%	11%	19%	69%
33	3	1%	11%	19%	69%
33	3	1%	11%	19%	69%
33	3	1%	11%	19%	69%
34	1	5%	44%	27%	24%
34	1	1%	55%	21%	23%
34	1	2%	41%	39%	19%
34	1	1%	56%	20%	23%
35	**1**	17%	41%	34%	9%
35	**1**	21%	36%	35%	7%
35	**2**	20%	27%	38%	15%
35	**2**	6%	25%	42%	27%
36	**0**	43%	19%	28%	10%
36	**1**	17%	39%	33%	11%
36	**1**	18%	37%	35%	11%
36	**2**	18%	34%	39%	9%
37	1	1%	75%	14%	11%
37	1	7%	37%	27%	29%
37	1	3%	57%	23%	17%
37	1	5%	49%	23%	23%
38	2	1%	8%	61%	31%
38	2	1%	8%	59%	32%
38	2	1%	7%	59%	33%
38	2	1%	10%	55%	35%
39	**2**	17%	19%	38%	26%
39	**0**	36%	23%	22%	19%
39	**0**	33%	23%	23%	21%
39	**0**	35%	26%	20%	19%
40	**1**	2%	41%	20%	37%
40	**1**	1%	39%	23%	37%
40	**3**	3%	31%	26%	41%
40	**1**	1%	44%	19%	36%
41	0	46%	21%	9%	24%
41	0	53%	21%	7%	19%
41	0	59%	19%	6%	17%

(conclusão)

Data	Agrupamento previsto	Probabilidade de a observação pertencer à classe:			
		0	1	2	3
42	**1**	1%	40%	27%	31%
42	**3**	0%	33%	32%	35%
42	**1**	0%	38%	37%	25%
42	**1**	0%	42%	34%	24%
43	**3**	6%	25%	21%	49%
43	**3**	4%	22%	27%	47%
43	**2**	5%	21%	37%	37%
43	**3**	6%	21%	36%	37%
44	0	67%	30%	1%	2%
44	0	77%	23%	0%	1%
44	0	76%	23%	0%	1%
44	0	71%	29%	0%	1%
45	**0**	34%	20%	23%	23%
45	**2**	22%	15%	35%	27%
45	**3**	9%	17%	27%	47%
45	**2**	29%	20%	30%	21%
46	2	1%	29%	51%	19%
46	2	2%	31%	49%	18%
46	2	1%	31%	49%	18%
46	2	2%	36%	41%	21%
47	0	51%	24%	18%	7%
47	0	51%	22%	21%	6%
47	0	45%	25%	24%	7%
47	0	51%	22%	21%	6%
48	3	1%	15%	19%	65%
48	3	2%	11%	18%	69%
48	3	1%	12%	10%	77%
49	**2**	17%	21%	39%	23%
49	**2**	19%	24%	33%	25%
49	**0**	51%	13%	16%	20%
49	**0**	49%	11%	21%	19%

Nota-se que, para os casos destacados em vermelho na segunda coluna da Tabela 10, existem diferentes classes direcionadas para a mesma data, porém para horários fiscais distintos. Isto constitui um comportamento esperado, pois, embora seja aferida uma única medida de TOG-G por dia, seu valor provavelmente se altera mediante as diferentes configurações das variáveis de processo ao longo do tempo.

Desta forma, visando auxiliar o processo de tomada de decisão ao longo do dia, extraíram-se as probabilidades de as observações da base fiscal pertencerem a cada um dos agrupamentos, também apresentadas na Tabela 10. Assim, torna-se possível à operação da plataforma avaliar o risco de desenquadramento a partir da probabilidade de determinada previsão pertencer à classe 3.

4.1.5. Etapa de Previsão - B

Passo B.1: considerando o conjunto de treinamento selecionado a partir da combinação com maior acurácia na etapa de classificação, indicado no Passo A.7, criou-se um modelo matemático para previsão do TOG-G, por meio de uma regressão linear múltipla, em função da informação de agrupamento, considerada como uma entrada categórica, em conjunto com as variáveis de processo mais significativas selecionadas na Etapa A.3.

Os modelos matemáticos gerados foram sumarizados em uma única equação que pode ser observada na Eq. (19), onde i indica o agrupamento, com $i = \{0,1,2,3\}$, e c_i o valor associado da constante. Neste caso, $c_0 = 3{,}93$, $c_1 = 10{,}90$, $c_2 = 16{,}84$ e $c_3 = 27{,}50$. Desta forma, para uma nova observação, a classificação do agrupamento por meio do algoritmo detalhado na Etapa de Classificação - A direciona à equação de previsão de TOG-G apropriada.

Os valores de ajuste obtidos com a regressão foram $R^2 = 94{,}25\%$, $R^2_{adj} = 93{,}90\%$ e $R^2_{pred} = 93{,}32\%$. É válido reforçar que, quanto mais precisa a indicação do agrupamento, mais correta a previsão do TOG-G.

$$\begin{aligned} TOGG_i = \; & c_i + 0{,}00311x_{101} + 0{,}02137x_{108} + 0{,}01232x_{122} \\ & - 0{,}002025x_{128} - 0{,}002546x_{161} + 0{,}00621x_{162} \\ & + 0{,}00563x_{168} + 0{,}1525x_{176} \end{aligned} \tag{19}$$

Para melhor visualização dos efeitos principais sobre o TOG-G, elaborou-se a Figura 17, que indica como cada uma das variáveis influencia na resposta de interesse.

Para que fosse possível avaliar o impacto de cada variável no modelo a partir da magnitude de seus coeficientes, optou-se pela padronização das variáveis, subtraindo a média e dividindo pelo desvio padrão. Os coeficientes associados à cada variável, bem como seu respectivo *p-value*, podem ser observados na Tabela 11.

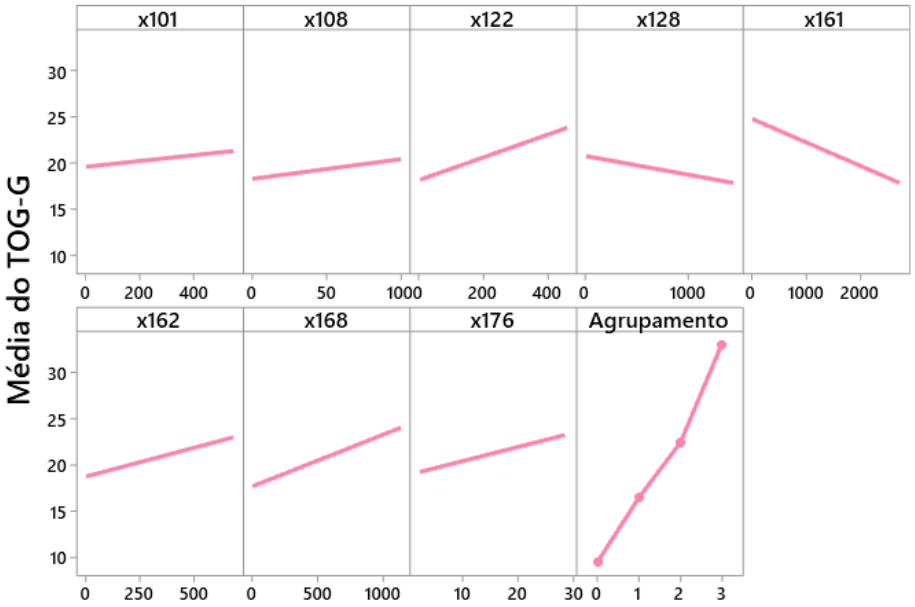


Figura 17 - Efeitos principais para a previsão do TOG-G

Tabela 11 - Coeficientes padronizados para os termos do modelo de previsão do TOG-G

Termo	Coeficiente padronizado	*p-value*
x_{101}	0,461	0,017
x_{108}	0,409	0,023
x_{122}	0,815	0,000
x_{128}	-0,817	0,001
x_{161}	-1,115	0,000
x_{162}	0,393	0,028
x_{168}	0,998	0,000
x_{176}	0,756	0,001

Neste caso, os valores para as constantes associadas a cada agrupamento i, sendo $i = \{0,1,2,3\}$, foram $c_{0_{norm}} = 9{,}439$, $c_{1_{norm}} = 16{,}410$, $c_{2_{norm}} = 22{,}354$ e $c_{3_{norm}} = 33{,}008$.

É possível perceber que x_{161} e x_{168} possuem os maiores coeficientes em módulo e, consequentemente, as maiores influências, confirmando a maior inclinação das retas relacionadas na Figura 17. A partir desta conclusão, os gráficos de linha de contorno (Figura 18) e de superfície (Figura 19), foram construídos direcionando x_{161} e x_{168} para os eixos x e y, respectivamente.

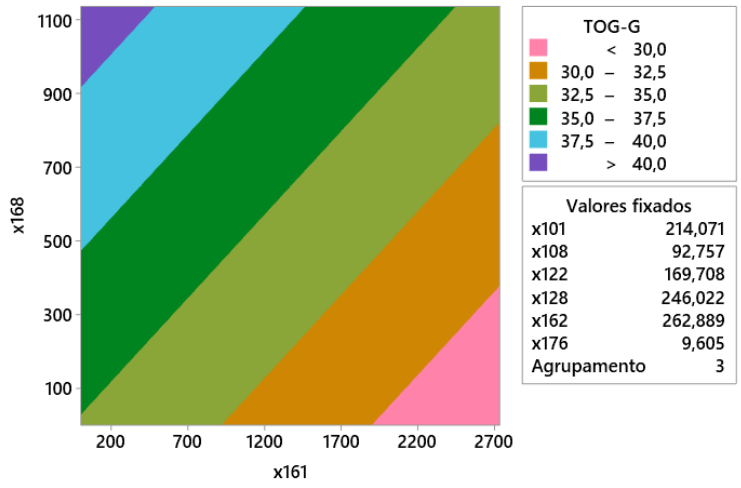


Figura 18 - Gráfico de linha de contorno para o modelo do TOG-G (Fonte: autoria própria)

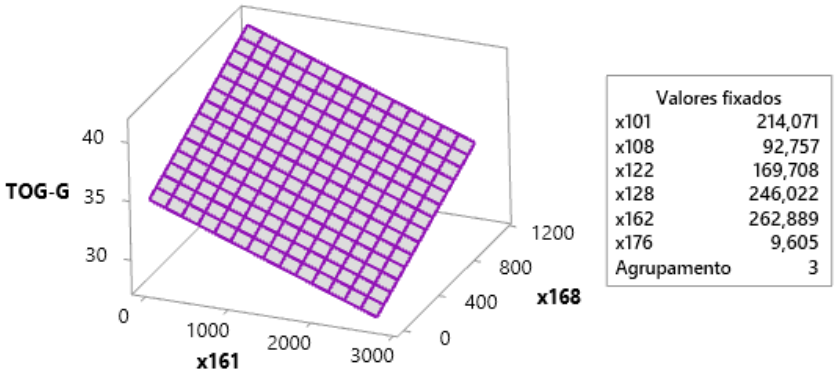


Figura 19 - Gráfico de superfície para o modelo do TOG-G (Fonte: autoria própria)

Passo B.2: aplicando o modelo de previsão de TOG-G ao conjunto de teste selecionado na etapa A.7, obtiveram-se os seguintes valores apresentados na Tabela 12, que apresenta também as métricas de MAE e MAPE calculadas com base nesses resultados. O coeficiente de determinação para este resultado foi de $R^2 = 56{,}04\%$. A Figura 20 apresenta as séries plotadas de TOG-G real e TOG-G predito para este conjunto de teste.

Tabela 12 - Valores de TOG-G previstos e métricas MAE e MAPE

Data	TOG-G Real	TOG-G Prev	$\lvert TOG\text{-}G_{Prev} - TOG\text{-}G_{Real} \rvert$	$\frac{\lvert TOG\text{-}G_{Prev} - TOG\text{-}G_{Real} \rvert}{TOG\text{-}G_{Real}}$
1	6	8,333	2,333	0,389
2	29	34,463	5,463	0,188
3	6	8,252	2,252	0,375
4	32	34,490	2,490	0,078
5	16	10,335	5,665	0,354

(conclusão)

Data	TOG-G Real	TOG-G Prev	$\lvert TOG\text{-}G_{Prev} - TOG\text{-}G_{Real} \rvert$	$\frac{\lvert TOG\text{-}G_{Prev} - TOG\text{-}G_{Real} \rvert}{TOG\text{-}G_{Real}}$
6	14	7,711	6,289	0,449
7	15	15,121	0,121	0,008
8	27	23,347	3,653	0,135
9	33	34,000	1,000	0,030
10	9	8,032	0,968	0,108
11	27	9,026	17,974	0,666
12	13	16,838	3,838	0,295
13	33	33,747	0,747	0,023
14	5	6,907	1,907	0,381
15	31	16,768	14,232	0,459
16	14	17,338	3,338	0,238
17	20	23,784	3,784	0,189
18	18	21,540	3,540	0,197
19	6	7,945	1,945	0,324
20	34	17,380	16,621	0,489
21	10	8,654	1,346	0,135
22	8	8,288	0,288	0,036
23	10	7,905	2,095	0,209
24	18	15,992	2,008	0,112
25	30	22,153	7,847	0,262
26	35	34,631	0,369	0,011
27	14	14,688	0,688	0,049
28	5	5,541	0,541	0,108
29	31	33,263	2,263	0,073
30	18	15,484	2,516	0,140
31	8	12,211	4,211	0,526
32	24	17,590	6,410	0,267
33	41	37,684	3,316	0,081
34	29	17,461	11,539	0,398
35	20	20,875	0,875	0,044
36	18	22,300	4,300	0,239
37	20	16,375	3,625	0,181
38	18	25,559	7,559	0,420
39	30	8,153	21,847	0,728
40	22	34,137	12,137	0,552
41	18	8,949	9,051	0,503
42	26	33,572	7,572	0,291
43	29	34,730	5,730	0,198
44	14	7,992	6,008	0,429
45	37	23,744	13,256	0,358
46	22	23,284	1,284	0,058
47	15	7,896	7,104	0,474
48	29	34,085	5,085	0,175
49	20	22,316	2,316	0,116
		MAE	5,129	-
		MAPE	-	0,256

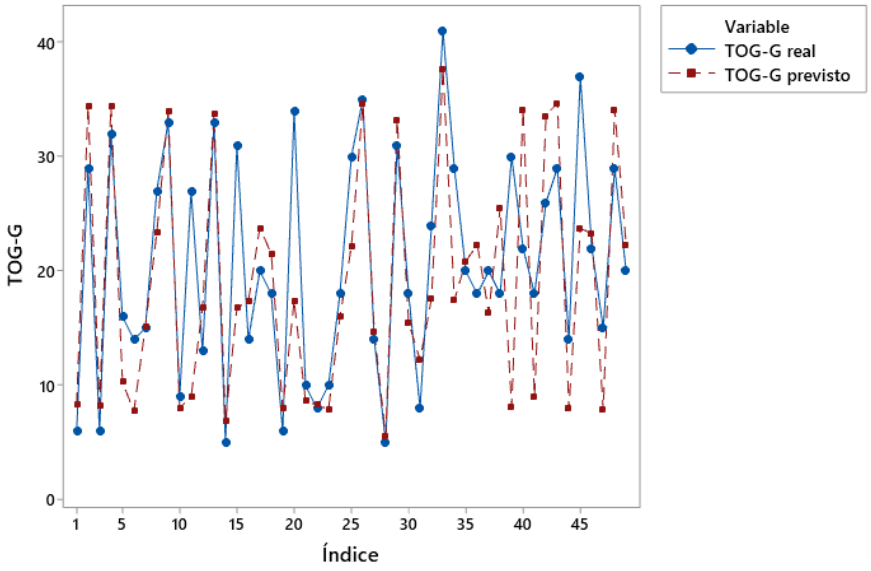


Figura 20 - Séries de TOG-G real e previsto para o conjunto de teste da base diária (Fonte: autoria própria)

Passo B.3: realizando a previsão do TOG-G para o conjunto de teste da base fiscal obtiveram-se os valores apresentados na Tabela 13.

Tabela 13 - Valores de TOG-G previstos para o conjunto de teste da base fiscal

Data	TOG-G Fiscal Previsto	Data	TOG-G Fiscal Previsto	Data	TOG-G Fiscal Previsto	Data	TOG-G Fiscal Previsto
1	6,067	13	32,471	25	31,307	37	16,454
1	6,759	13	31,490	25	21,086	38	24,377
1	6,923	13	31,581	26	32,666	38	24,309
1	6,933	14	5,226	26	32,895	38	23,915
2	32,889	14	5,343	26	33,587	38	22,984
2	22,665	14	5,226	26	32,725	39	19,807
2	22,551	14	5,181	27	6,036	39	6,173
2	31,790	15	16,310	27	13,033	39	6,508
3	6,594	15	16,733	27	13,029	39	6,388
3	6,581	15	17,095	27	13,034	40	14,570
3	6,591	15	16,934	28	4,497	40	14,566
3	6,593	16	17,046	28	4,704	40	41,136
4	31,194	16	17,541	28	2,567	40	16,480
4	31,514	16	17,441	28	3,745	41	6,961
4	30,352	16	17,324	29	33,446	41	7,344
4	30,440	17	23,578	29	16,547	41	7,417
5	16,457	17	23,810	29	33,281	41	7,423
5	15,388	17	23,683	29	33,179	42	15,607
5	15,245	17	24,066	30	14,950	42	31,598

(conclusão)

Data	TOG-G Fiscal Previsto	Data	TOG-G Fiscal Previsto	Data	TOG-G Fiscal Previsto	Data	TOG-G Fiscal Previsto
5	15,484	18	19,968	30	15,800	42	14,777
6	13,032	18	19,933	30	15,714	42	14,703
6	6,042	18	19,840	30	15,472	43	30,202
6	6,045	18	19,770	31	18,745	43	31,027
6	6,046	19	4,641	31	12,512	43	20,853
7	20,196	19	3,339	31	12,503	43	29,221
7	12,833	19	3,339	31	12,054	44	7,507
7	19,235	19	3,498	32	32,644	44	8,129
7	30,054	20	15,282	32	22,122	44	8,148
8	19,534	20	15,504	32	15,960	44	8,185
8	18,931	20	16,270	32	15,526	45	5,629
8	18,692	20	15,812	33	33,682	45	19,606
8	19,269	21	7,401	33	34,063	45	30,466
9	34,030	21	7,633	33	34,229	45	20,054
9	34,040	21	6,465	33	34,305	46	23,625
9	34,124	21	6,465	34	15,587	46	23,255
9	33,806	22	6,647	34	15,834	46	23,274
10	6,640	22	6,705	34	16,022	46	22,982
10	6,279	22	6,499	34	15,749	47	6,156
10	6,280	22	6,649	35	13,696	47	6,237
10	6,278	23	6,059	35	13,741	47	6,309
11	16,005	23	6,304	35	16,667	47	6,234
11	15,842	23	6,302	35	20,856	48	33,914
11	15,177	23	6,306	36	6,999	48	34,627
11	16,963	24	14,333	36	15,123	48	34,241
12	16,551	24	14,323	36	14,877	48	33,556
12	16,376	24	14,338	36	20,747	49	21,026
12	22,710	24	14,325	37	16,155	49	20,840
13	32,795	25	30,526	37	16,363	49	7,461

Passo B.4: É válido relembrar que a comparação imediata dos valores previstos para esta base não é possível de acontecer, pois apenas um valor de TOG-G real está disponível por data. Desta forma, extraíram-se os valores das médias e do terceiro quartil para o conjunto de 4 dados referentes a cada data. A Tabela 14 apresenta os resultados obtidos de TOG-G previstos calculados a partir das medidas de média e 3º quartil.

Tabela 14 - Média e 3º quartil dos valores de TOG-G previstos para a base fiscal

Data	TOG-G Real	TOG-G Fiscal Médio	TOG-G Fiscal 3º quartil
1	6	6,671	6,926
2	29	27,474	32,065
3	6	6,590	6,593
4	32	30,875	31,274
5	16	15,643	15,727

(conclusão)

Data	TOG-G Real	TOG-G Fiscal Médio	TOG-G Fiscal 3º quartil
6	14	7,791	7,793
7	15	20,580	22,661
8	27	19,106	19,335
9	33	34,000	34,061
10	9	6,369	6,370
11	27	15,997	16,245
12	13	19,810	22,934
13	33	32,084	32,552
14	5	5,244	5,255
15	31	16,768	16,974
16	14	17,338	17,466
17	20	23,784	23,873
18	18	19,878	19,942
19	6	3,704	3,783
20	34	15,717	15,926
21	10	6,991	7,459
22	8	6,625	6,662
23	10	6,243	6,304
24	18	14,330	14,334
25	30	25,818	30,722
26	35	32,968	33,068
27	14	11,283	13,033
28	5	3,878	4,549
29	31	29,113	33,322
30	18	15,484	15,735
31	8	13,953	14,070
32	24	21,563	24,752
33	41	34,070	34,247
34	29	15,798	15,881
35	20	16,240	17,715
36	18	14,437	16,529
37	20	16,375	16,472
38	18	23,896	24,326
39	30	9,719	9,833
40	22	21,688	22,644
41	18	7,286	7,418
42	26	19,171	19,605
43	29	27,826	30,408
44	14	7,992	8,157
45	37	18,939	22,657
46	22	23,284	23,362
47	15	6,234	6,255
48	29	34,084	34,337
49	20	14,195	20,886

A Figura 21 apresenta as séries plotadas de TOG-G real e de média e 3º quartil do TOG-G previsto, para este conjunto de teste da base fiscal.

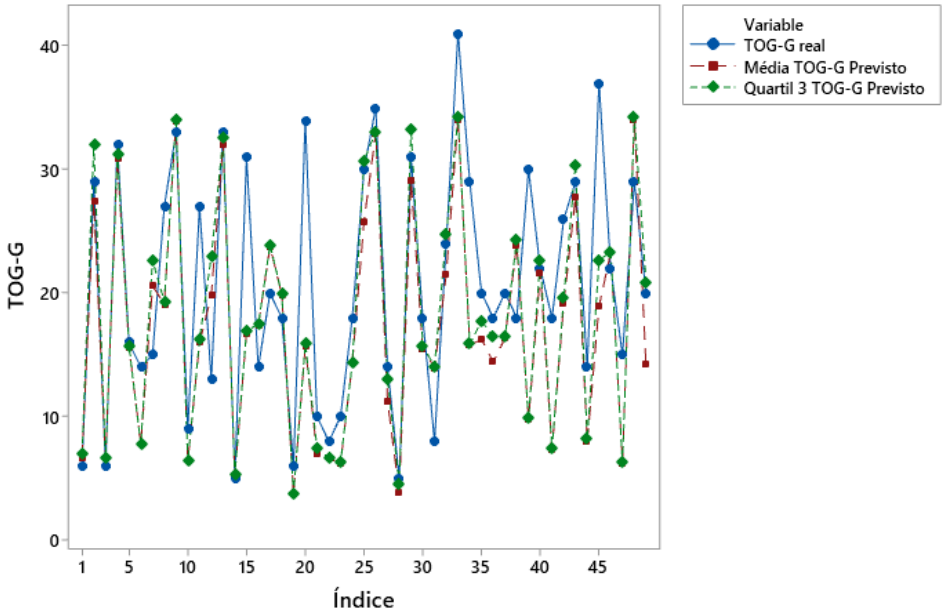


Figura 21 - Séries de TOG-G real, média e 3º quartil dos TOG-G previstos para o conjunto de teste da base fiscal (Fonte: autoria própria)

As métricas para os valores obtidos a partir da média e 3º quartil das previsões para o conjunto de teste da base fiscal, bem como para as previsões a partir do conjunto de teste da base diária podem ser observados na Tabela 15.

Tabela 15 - Métricas para os conjuntos de teste relativos às abordagens diária e fiscal e TOG-S

Métrica	TOG-G previsto		
	Base diária	Média (base fiscal)	3º quartil (base fiscal)
MAE	5,129	5,027	4,729
MAPE	0,256	0,255	0,243
R^2	56,04%	61,04%	60,43%
Correlação	0,749	0,781	0,777

4.1.6. Confirmação

Dado o conjunto de treinamento selecionado no Passo A.7, criou-se um modelo matemático para previsão do TOG-G, por meio de uma regressão linear múltipla, considerando apenas as variáveis de processo mais significativas indicadas na Etapa A. É importante reforçar que, nesta seção, a informação do agrupamento não foi alocada dentro do conjunto de variáveis preditoras.

O modelo matemático alcançado está expresso na Eq. (20), e seus valores de ajuste obtidos com a regressão foram $R^2 = 44{,}62\%$, $R^2_{adj} = 42{,}24\%$ e $R^2_{pred} = 37{,}96\%$.

$$TOGG_{SA} = 6{,}90 + 0{,}00564x_{101} + 0{,}0452x_{108} + 0{,}0176x_{122} - 0{,}008858x_{128} - 0{,}00169x_{161} - 0{,}00635x_{162} + 0{,}00943x_{168} + 0{,}735x_{176} \quad (20)$$

Aplicando o modelo de previsão ao conjunto de teste associado ao conjunto de treinamento previamente mencionado, obtiveram-se os seguintes valores apresentados na Tabela 16, que, por sua vez, também exibe as métricas de MAE e MAPE. Para este resultado, o coeficiente de determinação foi de $R^2 = 47{,}56\%$.

Tabela 16 - Valores de TOG-G previstos (sem agrupamento) e métricas MAE e MAPE

Data	TOG-G Real	TOG-G Prev	$\lvert TOG\text{-}G_{Prev} - TOG\text{-}G_{Real} \rvert$	$\frac{\lvert TOG\text{-}G_{Prev} - TOG\text{-}G_{Real} \rvert}{TOG\text{-}G_{Real}}$
1	6	13,682	7,682	1,280
2	29	29,287	0,287	0,010
3	6	11,942	5,942	0,990
4	32	25,322	6,678	0,209
5	16	20,902	4,902	0,306
6	14	7,752	6,248	0,446
7	15	18,597	3,597	0,240
8	27	25,159	1,841	0,068
9	33	27,134	5,866	0,178
10	9	10,891	1,891	0,210
11	27	21,885	5,115	0,189
12	13	24,478	11,478	0,883
13	33	26,409	6,591	0,200
14	5	7,298	2,298	0,460
15	31	23,712	7,288	0,235
16	14	22,958	8,958	0,640
17	20	26,126	6,126	0,306
18	18	17,438	0,562	0,031
19	6	10,734	4,734	0,789
20	34	24,082	9,918	0,292
21	10	13,930	3,930	0,393
22	8	16,820	8,820	1,102
23	10	12,071	2,071	0,207
24	18	14,783	3,217	0,179
25	30	21,152	8,848	0,295
26	35	27,884	7,116	0,203
27	14	9,774	4,226	0,302
28	5	4,961	0,039	0,008
29	31	23,503	7,497	0,242
30	18	22,070	4,070	0,226
31	8	11,703	3,703	0,463
32	24	24,565	0,565	0,024
33	41	33,651	7,349	0,179

(conclusão)

Data	TOG-G Real	TOG-G Prev	$\lvert TOG\text{-}G_{Prev} - TOG\text{-}G_{Real} \rvert$	$\frac{\lvert TOG\text{-}G_{Prev} - TOG\text{-}G_{Real} \rvert}{TOG\text{-}G_{Real}}$
34	29	24,302	4,698	0,162
35	20	17,400	2,600	0,130
36	18	19,136	1,136	0,063
37	20	22,371	2,371	0,119
38	18	37,018	19,018	1,057
39	30	17,298	12,702	0,423
40	22	38,432	16,432	0,747
41	18	16,480	1,520	0,084
42	26	24,468	1,532	0,059
43	29	26,600	2,400	0,083
44	14	4,316	9,684	0,692
45	37	19,677	17,323	0,468
46	22	26,723	4,723	0,215
47	15	12,622	2,378	0,159
48	29	28,149	0,851	0,029
49	20	20,827	0,827	0,041
		MAE	5,503	-
		MAPE	-	0,333

De forma semelhante, aplicou-se o mesmo modelo de previsão ao conjunto de teste da base fiscal, construído a partir do conjunto de teste da base diária, para o qual os seguintes resultados foram obtidos, conforme mostra a Tabela 17.

Tabela 17 - Valores de TOG-G previstos (sem agrupamento) para o conjunto de teste da base fiscal

Data	TOG-G Fiscal Previsto	Data	TOG-G Fiscal Previsto	Data	TOG-G Fiscal Previsto	Data	TOG-G Fiscal Previsto
1	6,067	13	32,471	25	31,307	37	16,454
1	6,759	13	31,490	25	21,086	38	24,377
1	6,923	13	31,581	26	32,666	38	24,309
1	6,933	14	5,226	26	32,895	38	23,915
2	32,889	14	5,343	26	33,587	38	22,984
2	22,665	14	5,226	26	32,725	39	19,807
2	22,551	14	5,181	27	6,036	39	6,173
2	31,790	15	16,310	27	13,033	39	6,508
3	6,594	15	16,733	27	13,029	39	6,388
3	6,581	15	17,095	27	13,034	40	14,570
3	6,591	15	16,934	28	4,497	40	14,566
3	6,593	16	17,046	28	4,704	40	41,136
4	31,194	16	17,541	28	2,567	40	16,480
4	31,514	16	17,441	28	3,745	41	6,961
4	30,352	16	17,324	29	33,446	41	7,344
4	30,440	17	23,578	29	16,547	41	7,417
5	16,457	17	23,810	29	33,281	41	7,423
5	15,388	17	23,683	29	33,179	42	15,607
5	15,245	17	24,066	30	14,950	42	31,598

(conclusão)

Data	TOG-G Fiscal Previsto	Data	TOG-G Fiscal Previsto	Data	TOG-G Fiscal Previsto	Data	TOG-G Fiscal Previsto
5	15,484	18	19,968	30	15,800	42	14,777
6	13,032	18	19,933	30	15,714	42	14,703
6	6,042	18	19,840	30	15,472	43	30,202
6	6,045	18	19,770	31	18,745	43	31,027
6	6,046	19	4,641	31	12,512	43	20,853
7	20,196	19	3,339	31	12,503	43	29,221
7	12,833	19	3,339	31	12,054	44	7,507
7	19,235	19	3,498	32	32,644	44	8,129
7	30,054	20	15,282	32	22,122	44	8,148
8	19,534	20	15,504	32	15,960	44	8,185
8	18,931	20	16,270	32	15,526	45	5,629
8	18,692	20	15,812	33	33,682	45	19,606
8	19,269	21	7,401	33	34,063	45	30,466
9	34,030	21	7,633	33	34,229	45	20,054
9	34,040	21	6,465	33	34,305	46	23,625
9	34,124	21	6,465	34	15,587	46	23,255
9	33,806	22	6,647	34	15,834	46	23,274
10	6,640	22	6,705	34	16,022	46	22,982
10	6,279	22	6,499	34	15,749	47	6,156
10	6,280	22	6,649	35	13,696	47	6,237
10	6,278	23	6,059	35	13,741	47	6,309
11	16,005	23	6,304	35	16,667	47	6,234
11	15,842	23	6,302	35	20,856	48	33,914
11	15,177	23	6,306	36	6,999	48	34,627
11	16,963	24	14,333	36	15,123	48	34,241
12	16,551	24	14,323	36	14,877	48	33,556
12	16,376	24	14,338	36	20,747	49	21,026
12	22,710	24	14,325	37	16,155	49	20,840
12	23,605	25	20,352	37	16,528	49	7,455
13	32,795	25	30,526	37	16,363	49	7,461

Seguindo a mesma estratégia aplicada no método, extraíram-se a média e o terceiro quartil para o conjunto de 4 previsões associadas a cada data, de forma que fosse possível calcular as métricas em função do valor de TOG-G real, cuja medida é diária. Os valores calculados estão apresentados na Tabela 18.

Tabela 18 - Média e 3º quartil dos valores de TOG-G previstos (sem agrupamento) para a base fiscal

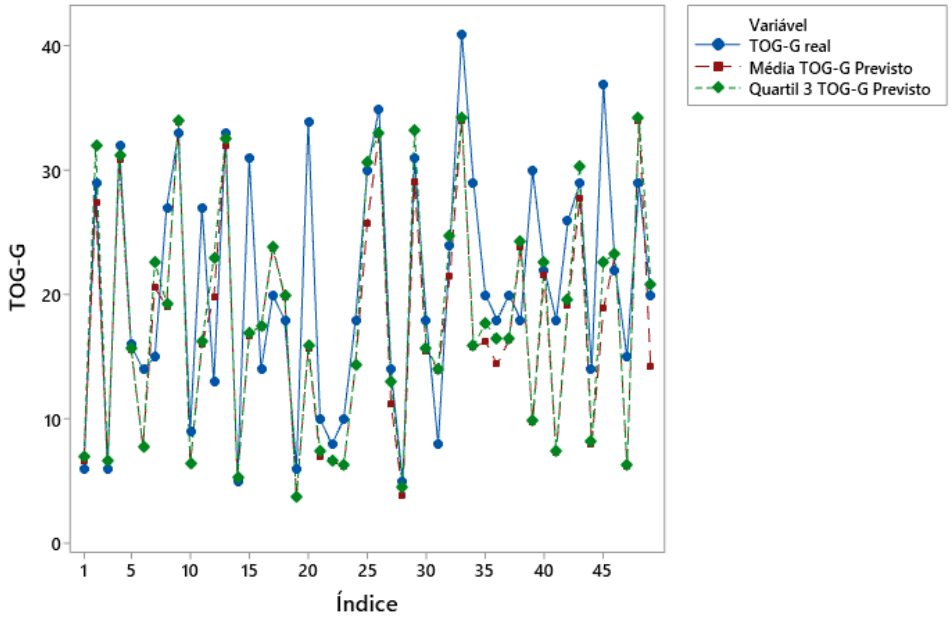
Data	TOG-G Real	TOG-G Fiscal Médio	TOG-G Fiscal 3º. quartil
1	6	22,897	25,063
2	29	38,432	60,187
3	6	22,070	23,314
4	32	20,342	21,732
5	16	20,298	21,536

(conclusão)

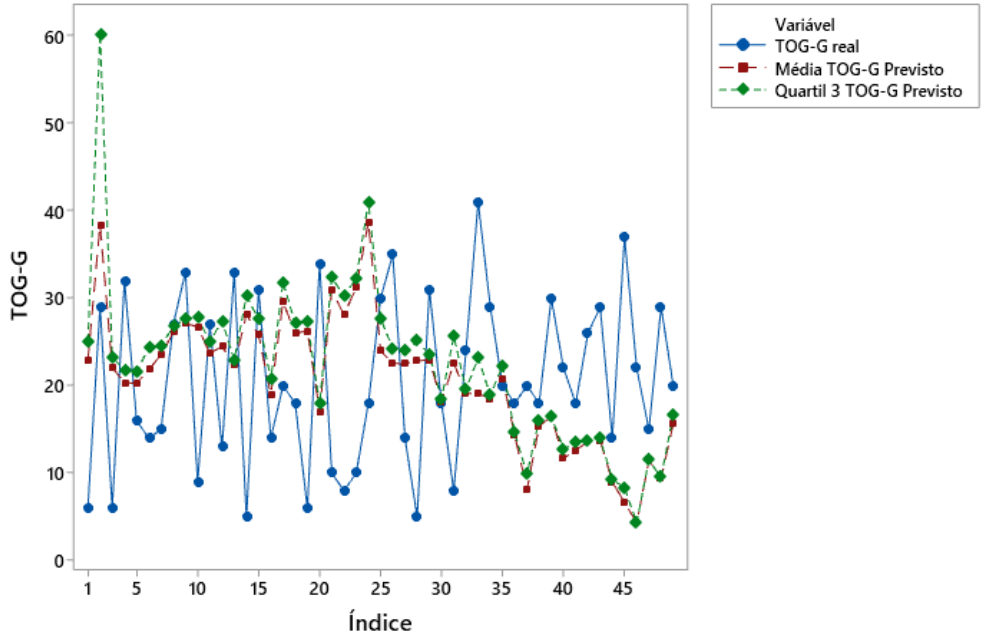
Data	TOG-G Real	TOG-G Fiscal Médio	TOG-G Fiscal 3º. quartil
6	14	21,885	24,451
7	15	23,503	24,566
8	27	26,126	26,783
9	33	27,134	27,628
10	9	26,723	27,881
11	27	23,712	25,006
12	13	24,478	27,420
13	33	22,371	22,951
14	5	28,149	30,288
15	31	25,783	27,594
16	14	18,999	20,770
17	20	29,585	31,838
18	18	26,003	27,130
19	6	26,266	27,308
20	34	17,058	17,890
21	10	30,988	32,413
22	8	28,110	30,225
23	10	31,227	32,289
24	18	38,719	41,033
25	30	23,981	27,696
26	35	22,541	24,263
27	14	22,528	24,061
28	5	22,853	25,172
29	31	22,958	23,537
30	18	18,181	18,517
31	8	22,603	25,658
32	24	19,139	19,569
33	41	19,101	23,190
34	29	18,521	18,899
35	20	20,837	22,310
36	18	14,323	14,596
37	20	8,116	9,891
38	18	15,383	16,049
39	30	16,484	16,497
40	22	11,703	12,768
41	18	12,592	13,468
42	26	13,643	13,650
43	29	13,772	13,965
44	14	8,999	9,327
45	37	6,662	8,302
46	22	4,316	4,369
47	15	11,475	11,492
48	29	9,453	9,496
49	20	15,631	16,725

A Figura 22(a) compila as séries plotadas de TOG-G real, média e 3º quartil do TOG-G previsto com agrupamento, provenientes do método proposto, e as mesmas séries plotadas com os valores previstos a partir da aplicação do modelo sem agrupamento

Figura 22(b), isto é, se fossem consideradas apenas as variáveis disponíveis dentro da plataforma.



(a)



(b)

Figura 22 - Séries de TOG-G real, média e 3º quartil dos TOG-G previstos (a) com agrupamento e (b) sem agrupamento para o conjunto de teste da base fiscal (Fonte: autoria própria)

Calcularam-se também as métricas MAE e MAPE, além do coeficiente de determinação R^2 e do coeficiente de correlação de Pearson, para os resultados gerados a partir da média e terceiro quartil das previsões da base fiscal e também para base de teste diária, sem agrupamento, apresentados mais à frente nesta seção.

Uma vez que o TOG-S é a medida disponibilizada alternativamente na plataforma para uso em tempo real, avaliaram-se as mesmas métricas considerando os valores reais de TOG-S e TOG-G. A Tabela 19 apresenta as medições de ambas variáveis para o conjunto de teste associado à base diária. É válido relembrar que os valores de TOG -S são sempre menores do que os valores de TOG-G, como pode ser mais facilmente observado na Figura 23, pois abrangem apenas a fração dispersa.

Tabela 19 - Valores de TOG-G e TOG-S reais e métricas MAE e MAPE

Data	TOG-G Real	TOG-S Real	$\lvert TOG\text{-}S_{Real} - TOG\text{-}G_{Real} \rvert$	$\frac{\lvert TOG\text{-}S_{Real} - TOG\text{-}G_{Real} \rvert}{TOG\text{-}G_{Real}}$
1	6	3,260	2,740	0,457
2	29	20,560	8,440	0,291
3	6	4,000	2,000	0,333
4	32	10,670	21,330	0,667
5	16	8,660	7,340	0,459
6	14	3,880	10,120	0,723
7	15	8,370	6,630	0,442
8	27	14,900	12,100	0,448
9	33	14,280	18,720	0,567
10	9	2,600	6,400	0,711
11	27	8,920	18,080	0,670
12	13	10,990	2,010	0,155
13	33	17,770	15,230	0,462
14	5	3,440	1,560	0,312
15	31	9,520	21,480	0,693
16	14	7,230	6,770	0,484
17	20	13,460	6,540	0,327
18	18	5,740	12,260	0,681
19	6	3,220	2,780	0,463
20	34	10,490	23,510	0,691
21	10	8,480	1,520	0,152
22	8	6,360	1,640	0,205
23	10	4,760	5,240	0,524
24	18	3,000	15,000	0,833
25	30	10,590	19,410	0,647
26	35	15,370	19,630	0,561
27	14	7,000	7,000	0,500
28	5	5,630	0,630	0,126
29	31	9,730	21,270	0,686
30	18	12,370	5,630	0,313
31	8	3,890	4,110	0,514

(conclusão)

Data	TOG-G Real	TOG-S Real	$\lvert TOG\text{-}S_{Real} - TOG\text{-}G_{Real} \rvert$	$\frac{\lvert TOG\text{-}S_{Real} - TOG\text{-}G_{Real} \rvert}{TOG\text{-}G_{Real}}$
32	24	12,260	11,740	0,489
33	41	18,020	22,980	0,560
34	29	10,600	18,400	0,634
35	20	8,260	11,740	0,587
36	18	9,750	8,250	0,458
37	20	8,010	11,990	0,600
38	18	31,240	13,240	0,736
39	30	7,990	22,010	0,734
40	22	31,800	9,800	0,445
41	18	3,860	14,140	0,786
42	26	13,020	12,980	0,499
43	29	16,580	12,420	0,428
44	14	2,000	12,000	0,857
45	37	7,180	29,820	0,806
46	22	13,550	8,450	0,384
47	15	3,510	11,490	0,766
48	29	14,560	14,440	0,498
49	20	10,290	9,710	0,486
		MAE	11,484	-
		MAPE	-	0,528

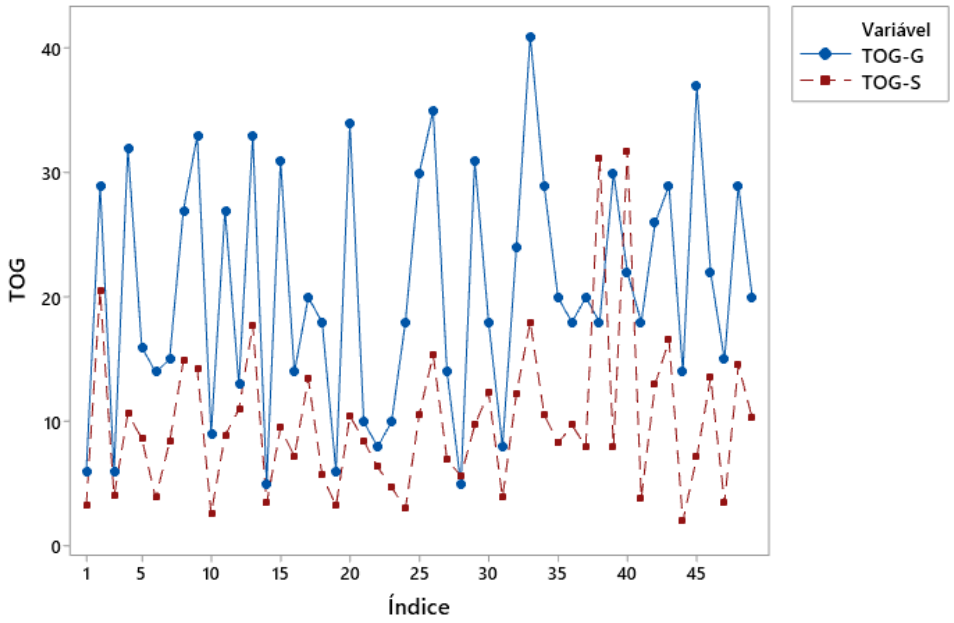


Figura 23 - Séries de TOG-G real e TOG-S real para o conjunto de teste da base diária (Fonte: autoria própria)

As Figura 24, Figura 25 e Figura 26 apresentam o *matrix plot* para os dados do TOG-G diário real e dos resultados obtidos com o método proposto, respectivamente: TOG-G diário previsto, média do TOG-G fiscal previsto e 3º quartil do TOG-G fiscal previsto. A

mesma representação para as medições diárias reais de TOG-S pode ser observada na Figura 27.

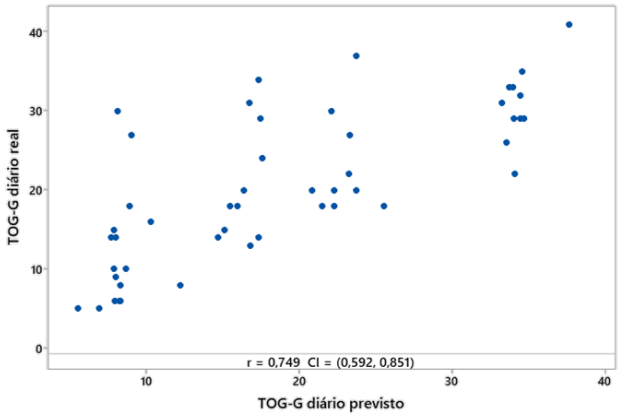


Figura 24 - *Matrix plot* do TOG-G diário real e TOG-G diário previsto (Fonte: autoria própria)

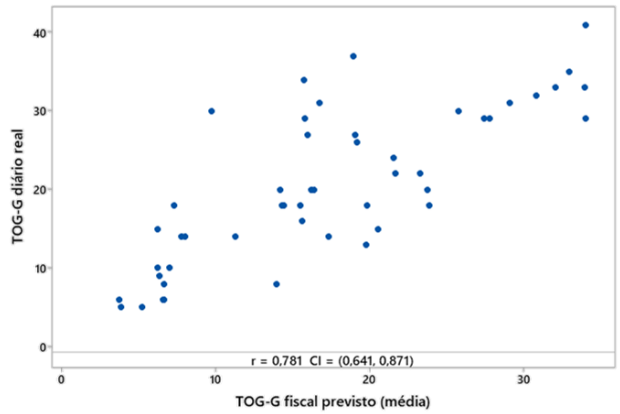


Figura 25 - *Matrix plot* do TOG-G diário real e média do TOG-G fiscal previsto (Fonte: autoria própria)

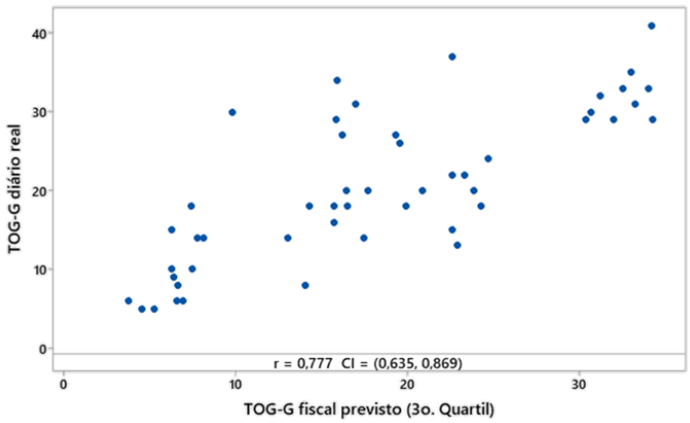


Figura 26 - *Matrix plot* do TOG-G diário real e 3o. quartil do TOG-G fiscal previsto (Fonte: autoria própria)

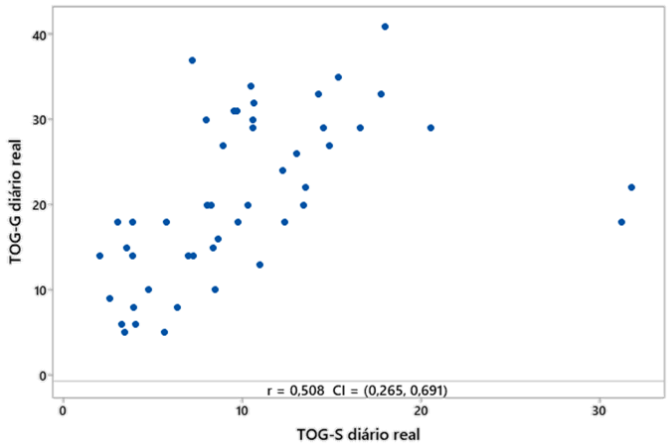


Figura 27 - *Matrix plot* do TOG-G diário real e TOG-S diário real (Fonte: autoria própria)

Revisitando os resultados obtidos através da aplicação do método proposto, tanto as métricas para o modelo sem agrupamento quanto para a comparação TOG-G e TOG-S foram agregadas ao conteúdo da Tabela 15, gerando a Tabela 20, de maneira a facilitar a visualização dos ganhos alcançados com o método proposto tanto em relação à regressão construída apenas com variáveis oriundas da plataforma quanto em relação à medida tida como referência dentro da plataforma hoje, o TOG-S.

Tabela 20 - Métricas para os conjuntos de teste relativos às abordagens diária e fiscal (com e sem agrupamento) e TOG-S

Métrica	TOG-S diário real	TOG-G previsto					
		Base diária		Média (base fiscal)		3º quartil (base fiscal)	
		C.A.*	S.A.*	C.A.	S.A.	C.A.	S.A.
MAE	11,484	5,129	5,503	5,027	11,368	4,729	12,022
MAPE	0,528	0,256	0,333	0,255	0,881	0,243	0,956
R^2	25,82%	56,04%	47,56%	61,04%	5,08%	60,43%	2,43%
Correlação	0,508	0,749	0,690	0,781	-0,225	0,777	-0,155

***C.A. = com agrupamento (método proposto) e S.A. = sem agrupamento**

Os *p-values* associados às correlações dos pares TOG-G diário real - Média (base fiscal) S.A. e TOG-G diário real - 3º quartil (base fiscal) S.A. foram, respectivamente, 0,120 e 0,285, o que mostra que estas correlações não são significativas. Para as demais correlações, o *p-value* atingiu o valor de 0,000, comprovando que elas são estatisticamente significativas, o que satisfaz a determinação do órgão regulamentador CONAMA no que diz respeito à utilização de novos métodos de controle do TOG-G.

Ainda pela observação da Tabela 20, é possível depreender que os dados de TOG-S, usados atualmente pela plataforma para controle dos níveis de TOG na água produzida, apresentam a menor correlação estatisticamente significativa com o TOG-G, quando comparado às previsões alcançadas com o método proposto. Dentre todos os resultados, aqueles obtidos a partir da média dos 4 valores previstos para os horários fiscais indicariam maior adequação para utilização *offshore*.

4.2. Conjunto clássico de dados

4.2.1. Considerações iniciais

Optou-se por realizar uma etapa adicional de aplicação do método proposto a um conjunto clássico de dados de regressão linear, isto é, dados que conhecidamente geram um modelo matemático com alto nível de ajuste, visando complementar a demonstração dos ganhos com o método proposto. Para este problema, a partir de um conjunto de variáveis preditoras envolvendo medidas diversas de diferentes espécies de peixes, espera-se prever seu peso em gramas.

Devido à dificuldade em encontrar uma base de dados com as mesmas características do TOG-G, cuja variável de resposta esteja associada a quatro amostras, as etapas do método relacionadas a esta particularidade serão suprimidas nesta subseção, sem prejuízos à demonstração do método.

4.2.2 Pré-processamento

O conjunto de dados original possuía 159 registros contendo informações relacionadas a 6 variáveis de entrada, além da variável de resposta, conforme descrito no Quadro 4.

Quadro 4 - Composição da base de dados clássica

Variável	Unidade de medida	Descrição	Tipo do dado
Comprimento 1	cm	Comprimento vertical	Real
Comprimento 2	cm	Comprimento diagonal	Real
Comprimento 3	cm	Comprimento cruzado	Real
Altura	cm	Altura	Real
Largura	cm	Largura	Real
Espécie	-	Espécie do peixe	Categórico
Peso	g	Peso	Real

A compilação da base foi realizada por meio de programação computacional utilizando a linguagem Python.

4.2.3. Pré-análise

Realizou-se uma análise de consistência de valores dentro da base, considerando a relevância deste aspecto para a construção do modelo preditivo, e não se detectaram a presença de valores faltantes. Entretanto, um valor inconsistente, igual a 0, para a variável Peso foi identificado, de maneira que se fez necessária a exclusão deste registro para não prejudicar a geração do modelo.

Na sequência, avaliaram-se as correlações existentes entre as variáveis preditoras e a variável de interesse. A Tabela 21 apresenta os valores de correlação e seus intervalos de confiança associados, bem como seus respectivos *p-values*. Todas as correlações com a variável de resposta demonstraram ser superior a 0,700 e foram apontadas como significativas ($p\text{-}value < 0{,}05$).

Tabela 21 - Correlação de Pearson entre as variáveis preditoras da base de dados clássica e a variável Peso

Variável	Correlação	95% IC para ρ	*p-value*
Comprimento 1	0,916	(0,886;0,938)	0,000
Comprimento 2	0,919	(0,890;0,940)	0,000
Comprimento 3	0,923	(0,896;0,943)	0,000
Altura	0,724	(0,641;0,791)	0,000
Largura	0,827	(0,848;0,916)	0,000

Optou-se por desconsiderar a variável categórica presente na base de dados e trabalhar apenas com as variáveis numéricas no conjunto de previsores, considerando a característica de classificação presente no método.

Para a aplicação do método proposto, portanto, considerou-se a base de dados original, subtraída da variável Espécie, contendo as 5 variáveis de entrada elencadas nesta subseção, além da variável de interesse Peso.

4.2.4. Etapa de Classificação - A

Passo A.1: para definição dos agrupamentos, plotou-se um *boxplot* para os dados do Peso, obtendo-se os valores do primeiro quartil, mediana e terceiro quartil, conforme indicado na Figura 28.

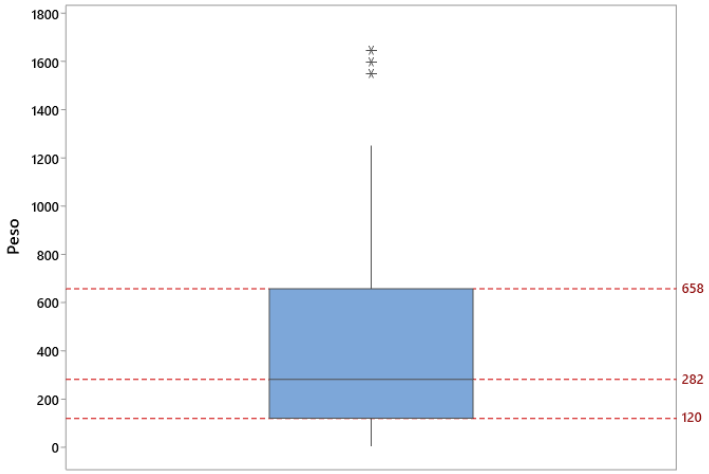


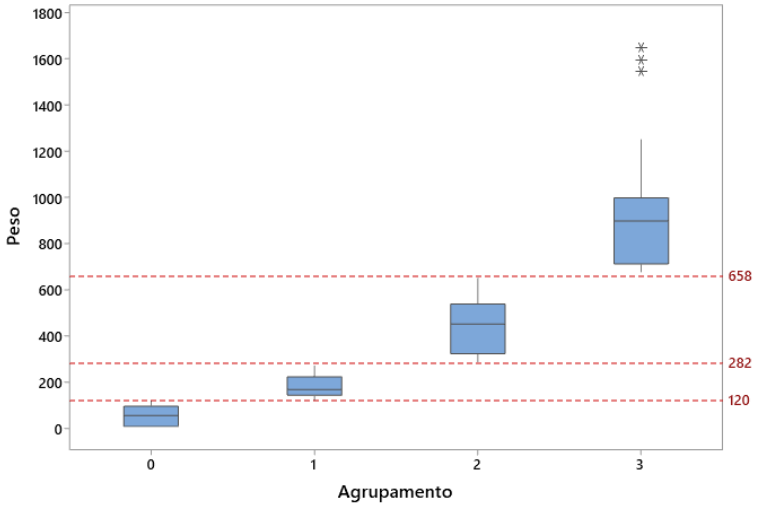
Figura 28 - *Boxplot* do Peso (Fonte: autoria própria)

Dada a inexistência de uma classe de interesse específica para este problema, encoraja-se a utilização dos valores identificados como limites para os agrupamentos de dados. Segundo a estatística descritiva, o quartil corresponde a um dos 3 valores que subdivide em 4 partes iguais qualquer conjunto ordenado de dados, ou seja, a aplicação de determinados valores como limite para agrupamentos de dados automaticamente produz um conjunto de dados balanceado em função da classe. Os limites inferior e superior para cada agrupamento, e o total de registros resultantes para cada um deles podem ser observados na Tabela 22.

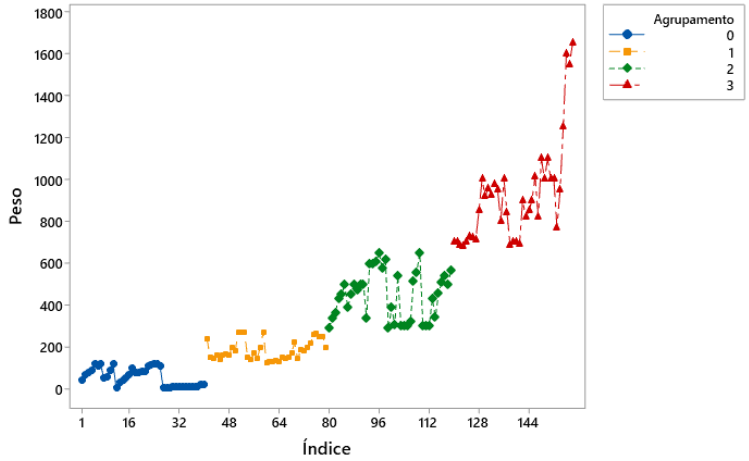
Tabela 22 - Limites inferior e superior e quantidade de registros para cada agrupamento

Agrupamento	Limite inferior	Limite superior	Quantidade de registros
0	0	120	40
1	121	282	39
2	283	658	40
3	659	1650	39

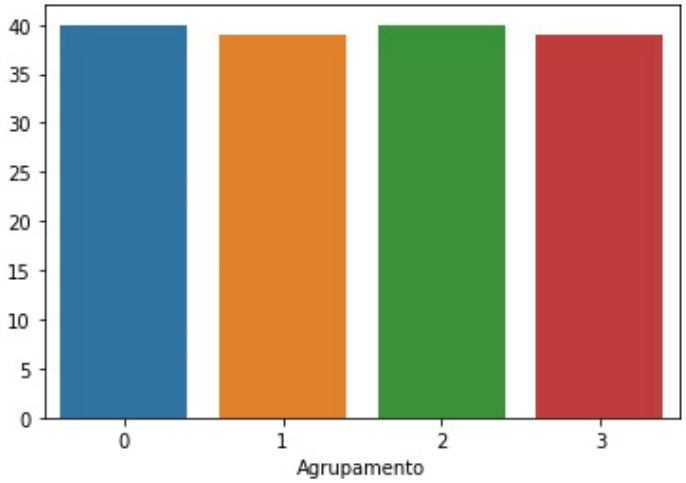
Passo A.2: conforme mencionado anteriormente, definir os valores dos quartis como os limites para os agrupamentos de dados, já gerou uma base balanceada em função destes agrupamentos, conforme reforçado nas Figura 29(a), Figura 29(b) e Figura 29(c).



(a)



(b)



(c)

Figura 29 - *Boxplot* (a), *time series plot* (b) e gráfico de barras (c) do Peso para o conjunto de dados balanceado (Fonte: autoria própria)

Passo A.3: a partir da base de dados balanceada, cuja composição é idêntica à base original, resolveu-se um problema de regressão linear múltipla, considerando as 5 variáveis preditoras originais, o Peso e o agrupamento. Devido à não intenção de redução do número de variáveis, não foi aplicado nenhum método de *stepwise*. A Tabela 23 apresenta a análise de variância (ANOVA) relativa à regressão.

Tabela 23 - ANOVA do modelo de regressão linear do Peso

Fonte de variação	GL	SQ	SQA	*F-value*	*p-value*
Regressão	8	19020631	2377579	331,96	0,000
Comprimento 1	1	18464	18464	2,58	0,110
Comprimento 2	1	2105	2105	0,29	0,589
Comprimento 3	1	1867	1867	0,26	0,610
Altura	1	69496	69496	9,7	0,002
Largura	1	3026	3026	0,42	0,517
Agrupamento	3	1233264	411088	57,4	0,000
Erro	149	1067160	7162		
Lack-of-fit	148	1065910	7202	5,76	0,322
Erro puro	1	1250	1250		
Total	157	20087791			

Passo A.4: etapa suprimida.

Passo A.5: realizou-se a divisão da base de dados em 50 pares distintos (treinamento e teste), na proporção de 80/20, respectivamente, por meio de programação computacional utilizando a função de validação cruzada *Stratified K-Folds* da linguagem Python.

Passo A.6: Aplicaram-se os algoritmos *Random Forest Classifier*, *K-Nearest Neighbors*, *Multilayer Perceptron*, Regressão Logística e *Support Vector Machine*, através de programação computacional utilizando a linguagem Python, para o desenvolvimento do classificador de agrupamentos a partir do conjunto de variáveis de entrada determinado no Passo A.3.

A Tabela 24 lista os parâmetros para os cinco algoritmos aplicados aos 50 conjuntos de treinamento extraídos, e suas respectivas acurácias obtidas. Para garantir a reprodutibilidade, o $random\ state$ foi mantido em 0 durante a execução do código.

Passo A.7: o algoritmo de Regressão Logística apresentou a melhor acurácia, sob uma porcentagem de 86,92%, de maneira que o tornou o algoritmo elencado para construir o classificador a ser utilizado para este problema.

Tabela 24 - Parâmetros e acurácias dos algoritmos de classificação

Algoritmo	Parâmetros	Acurácia
RF	$n_{estimators} = 150$ $max_{features} = \sqrt{n_{features}} = 3$	86,10%
KNN	$n_{neighbors} = 5$ $p = 2$	83,34%
MLP	$hiddenlayer_{sizes} = (50,50,50)$ $activation = 'relu'$ $solver = 'adam'$ $alpha = 0{,}0001$ $learning_{rate} = 'invscaling'$	71,46%
RL	$max_{iter} = 1000$	86,92%
SVM	$kernel = 'rbf'$ $C = 2$	81,23%

A combinação de conjuntos de treinamento e teste associados à maior acurácia individual gerada pelo algoritmo selecionado foi extraída, de forma que a classificação das observações do conjunto de teste fosse realizada. A Tabela 25 apresenta a proporção de dados por agrupamento para o par treinamento-teste selecionado.

Tabela 25 - Distribuição dos agrupamentos dentro dos conjuntos de treinamento e conjunto de teste

Agrupamento	Conjunto de Treinamento	Conjunto de teste
0	35	5
1	31	8
2	30	10
3	30	9

Aplicando o classificador às observações do conjunto de teste, obtiveram-se os valores de agrupamento listados na Tabela 26.

Passo A.8: para este resultado, foram calculadas as métricas de precisão, *recall* e *f1-score*, descritas na Tabela 27, bem como a matriz de confusão, apresentada na Tabela 28.

Tabela 26 - Classificações de agrupamento para o conjunto de teste

Observação	Agrupamento real	Agrupamento previsto
1	2	2
2	2	2
3	0	0
4	2	2
5	3	3
6	2	2
7	0	0
8	3	3
9	3	3
10	2	2
11	3	3
12	3	3
13	2	2
14	0	1
15	1	1
16	1	1
17	2	2
18	3	3
19	3	3
20	1	1
21	0	1
22	0	0
23	1	1
24	1	1
25	3	3
26	2	2
27	1	1
28	1	1
29	1	1
30	2	2
31	3	3
32	2	2

Tabela 27 - Métricas de avaliação para o classificador

Agrupamento	Precisão	*Recall*	*f-1 score*
0	1,00	0,60	0,75
1	0,80	1,00	0,89
2	1,00	1,00	1,00
3	1,00	1,00	1,00
Acurácia			0,94

Tabela 28 - Matriz de confusão para a previsão do Agrupamento do conjunto de testes da base

		Agrupamento Previsto				
		0	**1**	**2**	**3**	**Total**
Agrupamento Real	**0**	3	2	0	0	5
	1	0	8	0	0	8
	2	0	0	10	0	10
	3	0	0	0	9	9
	Total	3	10	10	9	32

Passo A.9: etapa suprimida.

4.2.5. Etapa de Previsão - B

Passo B.1: a partir do par treinamento-teste associado à maior acurácia do classificador no Passo A.7, criou-se um modelo matemático para previsão do Peso, considerando apenas o conjunto de treinamento. Aplicou-se a regressão linear múltipla, agregando a informação do agrupamento às variáveis preditoras da base original indicadas na Etapa A.3.

Geraram-se 4 modelos matemáticos, cada qual referente a um agrupamento, que podem ser observados nas equações de Eq. (21) a Eq. (24). Assim, a cada nova observação, o algoritmo determina seu agrupamento por meio do modelo de classificação e direciona esta informação, juntamente dos dados originais, à equação de previsão do Peso adequada. Isto reforça a importância da acurácia do modelo de classificação dentro do método proposto para a previsão do Peso.

$$Peso_0 = -390{,}1 + 91{,}2Comprimento_1 - 54{,}9Comprimento_2 - 8{,}2Comprimento_3 + 22{,}08Altura + 10{,}8Largura \quad (21)$$

$$Peso_1 = -503{,}1 + 91{,}2Comprimento_1 - 54{,}9Comprimento_2 - 8{,}2Comprimento_3 + 22{,}08Altura + 10{,}8Largura \quad (22)$$

$$Peso_2 = -510{,}9 + 91{,}2Comprimento_1 - 54{,}9Comprimento_2 - 8{,}2Comprimento_3 + 22{,}08Altura + 10{,}8Largura \quad (23)$$

$$Peso_3 = -291{,}6 + 91{,}2Comprimento_1 - 54{,}9Comprimento_2 - 8{,}2Comprimento_3 + 22{,}08Altura + 10{,}8Largura \quad (24)$$

A regressão apresentou valores de ajuste de $R^2 = 95{,}19\%$, $R^2_{adj} = 94{,}86\%$ e $R^2_{pred} = 93{,}77\%$.

Os efeitos principais das variáveis preditoras sobre o Peso podem ser observados na Figura 30, que indica como cada uma das variáveis influencia na resposta de interesse.

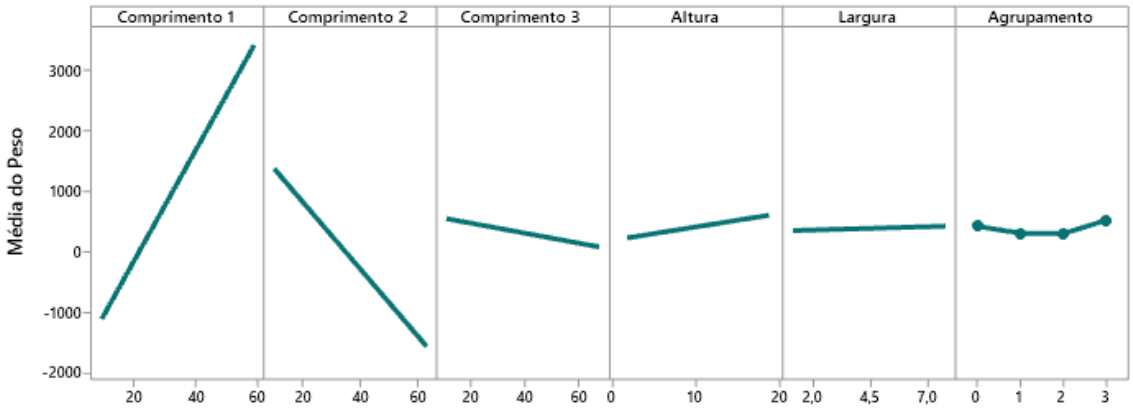


Figura 30 - Efeitos principais para a previsão do Peso

Padronizando-se as variáveis, é possível quantificar o impacto de cada uma delas no modelo de previsão do Peso mediante a observação da magnitude de seus coeficientes. Neste caso, optou-se pela subtração da média e divisão pelo desvio padrão para realizar a padronização. A Tabela 29 apresenta os coeficientes e respectivos *p-value* associados à cada variável, além dos valores para as constantes do modelo de cada agrupamento.

Tabela 29 - Coeficientes padronizados para os termos do modelo de previsão do Peso

Termo	Coeficiente padronizado	*p-value*
Comprimento 1	894,0	0,009
Comprimento 2	-579,0	0,113
Comprimento 3	-94,0	0,533
Altura	95,2	0,001
Largura	18,4	0,508
Agrupamento		
0	418,9	0,000
1	-113,1	0,000
2	-120,8	0,007
3	98,5	0,106

É fácil verificar que a variável Comprimento 1 possui o maior coeficiente em módulo e, portanto, a maior influência, seguido do Comprimento 2. De posse dessas informações, construíram-se os gráficos de linha de contorno, (Figura 31) e de superfície (Figura 32), direcionando o Comprimento 1 e o Comprimento 2 para os eixos x e y, respectivamente.

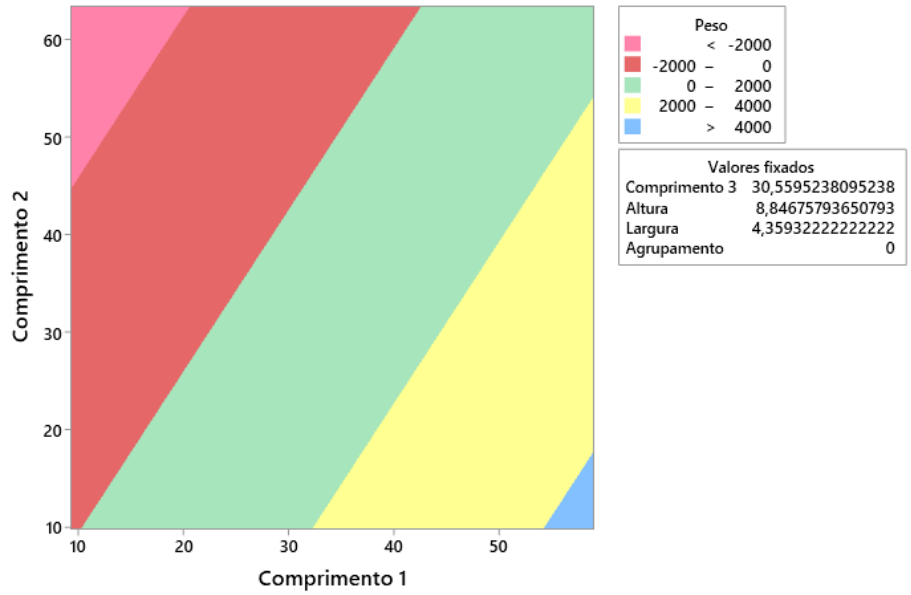


Figura 31 - Gráfico de linha de contorno para o modelo do Peso (Fonte: autoria própria)

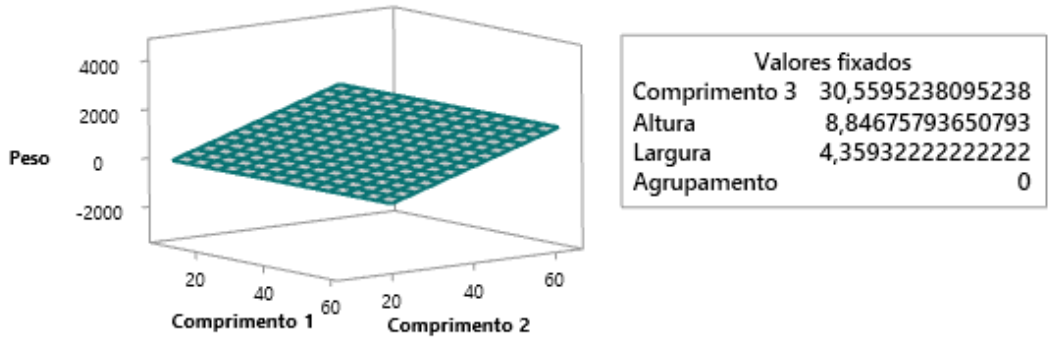


Figura 32 - Gráfico de superfície para o modelo do Peso (Fonte: autoria própria)

Passo B.2: realizando a previsão do Peso para o conjunto de teste indicado na etapa A.7, obtiveram-se os valores apresentados na Tabela 30. A mesma tabela apresenta as métricas de MAE e MAPE calculadas com base nesses resultados. O coeficiente de determinação e a correlação com os dados reais alcançados com este resultado foram de $R^2 = 91{,}90$ e

$\rho = 0{,}959$. A Figura 33 apresenta as séries plotadas de Peso real e Peso previsto (com agrupamento) para este conjunto de teste.

Tabela 30 - Valores de Peso previstos e métricas MAE e MAPE

Peso Real	**Peso Previsto**	$\lvert Peso_{Prev} - Peso_{Real} \rvert$	$\frac{\lvert Peso_{Prev} - Peso_{Real} \rvert}{Peso_{Real}}$
500,00	466,502	33,498	0,067
363,00	350,376	12,624	0,035
500,00	390,267	109,733	0,219
1015,00	893,386	121,614	0,120
456,00	638,453	182,453	0,400
1550,00	1275,501	274,499	0,177
1000,00	873,275	126,725	0,127
500,00	662,248	162,248	0,324
1000,00	1017,107	17,107	0,017
700,00	826,057	126,057	0,180
567,00	729,131	162,131	0,286
120,00	92,253	27,747	0,231
160,00	120,538	39,462	0,247
161,00	191,807	30,807	0,191
450,00	372,092	77,908	0,173
700,00	783,410	83,410	0,119
680,00	762,284	82,284	0,121
270,00	217,604	52,396	0,194
120,00	89,700	30,300	0,252
218,00	295,635	77,635	0,356
150,00	115,929	34,071	0,227
725,00	778,681	53,681	0,074
430,00	519,046	89,046	0,207
140,00	127,294	12,706	0,091
170,00	141,483	28,517	0,168
145,00	110,867	34,133	0,235
340,00	283,246	56,754	0,167
1000,00	975,847	24,153	0,024
540,00	637,563	97,563	0,181
	MAE	77,974	-
	MAPE	-	0,180

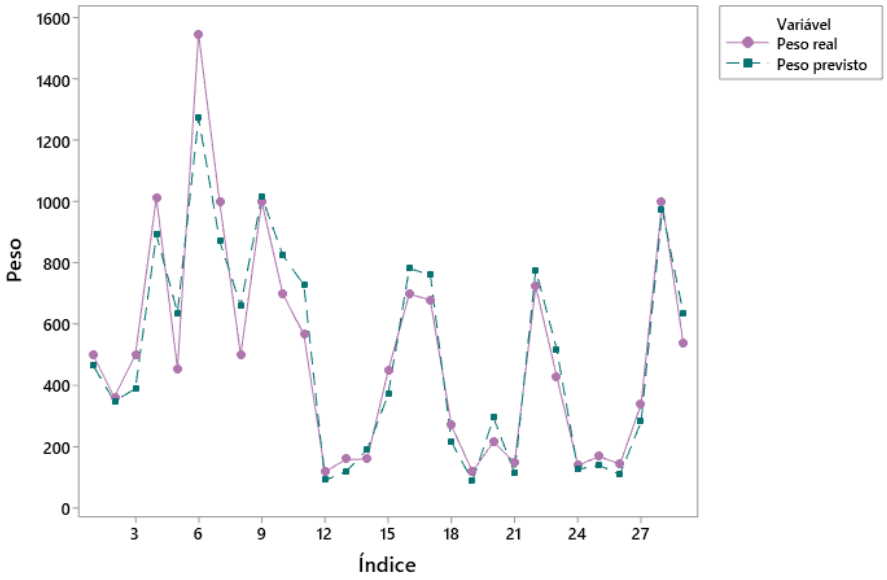


Figura 33 - Séries de Peso real e previsto (com agrupamento) para o conjunto de teste (Fonte: autoria própria)

Passo B.3: etapa suprimida.

Passo B.4: etapa suprimida.

4.2.6. Confirmação

Construiu-se uma regressão linear múltipla a partir do mesmo conjunto de treinamento indicado na etapa A.7, considerando apenas as 5 variáveis originalmente presentes e desconsiderando a informação do agrupamento. Na sequência, realizou-se a previsão do Peso e calcularam-se as mesmas métricas definidas para o método. A Tabela 31 apresenta os valores de Peso real e previsto (sem agrupamento) para o conjunto de teste associado. A Figura 34 apresenta as séries plotadas de Peso real e Peso previsto (sem agrupamento) para este conjunto de teste.

Tabela 31 - Valores de Peso real e previsto (sem agrupamento) e métricas MAE e MAPE

Peso Real	**Peso Previsto**	$\lvert Peso_{Prev} - Peso_{Real} \rvert$	$\frac{\lvert Peso_{Prev} - Peso_{Real} \rvert}{Peso_{Real}}$
500,00	554,630	54,630	0,109
363,00	421,415	58,415	0,161
500,00	461,478	38,522	0,077
1015,00	821,292	193,708	0,191
456,00	762,577	306,577	0,672

(conclusão)

Peso Real	**Peso Previsto**	$\lvert Peso_{Real} - Peso_{Prev} \rvert$	$\frac{\lvert Peso_{Real} - Peso_{Prev} \rvert}{Peso_{Real}}$
1550,00	1213,001	336,999	0,217
1000,00	761,243	238,757	0,239
500,00	781,112	281,112	0,562
1000,00	960,100	39,900	0,040
700,00	747,347	47,347	0,068
567,00	871,794	304,794	0,538
120,00	159,936	39,936	0,333
160,00	164,001	4,001	0,025
161,00	239,156	78,156	0,485
450,00	434,923	15,077	0,034
700,00	654,133	45,867	0,066
680,00	618,561	61,439	0,090
270,00	280,697	10,697	0,040
120,00	131,821	11,821	0,099
218,00	399,380	181,380	0,832
150,00	157,244	7,244	0,048
725,00	635,910	89,090	0,123
430,00	634,488	204,488	0,476
140,00	192,944	52,944	0,378
170,00	215,634	45,634	0,268
145,00	179,059	34,059	0,235
340,00	338,305	1,695	0,005
1000,00	908,944	91,056	0,091
540,00	760,892	220,892	0,409
	MAE	106,767	-
	MAPE	-	0,238

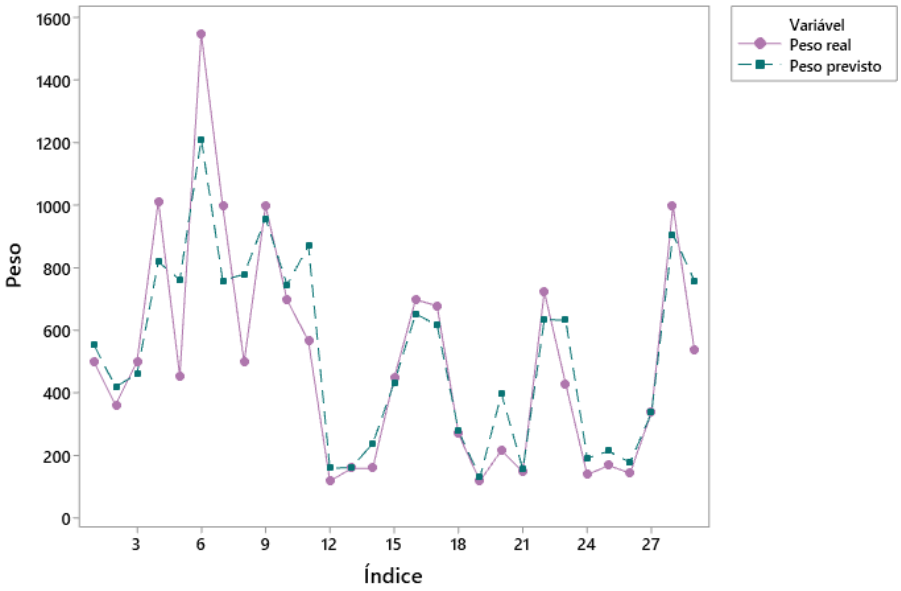


Figura 34 - Séries de Peso real e previsto (sem agrupamento) para o conjunto de teste (Fonte: autoria própria)

As métricas MAE e MAPE, além do coeficiente de determinação R^2, para o Peso previsto (sem agrupamento) foram calculados e estão apresentados na Tabela 32, que também revisita os resultados obtidos através da aplicação do método proposto neste conjunto de dados clássico.

Tabela 32 - Métricas para o conjunto de teste relativos às abordagens com agrupamento e sem agrupamento

Métrica	Peso previsto (com agrupamento)	Peso previsto (sem agrupamento)
MAE	77,974	106,767
MAPE	0,180	0,238
R^2	91,90%	82,56%

As Figura 35 e Figura 36 apresentam o *matrix plot* para as informações de Peso real e Peso previsto (sem agrupamento) e de Peso real com as previsões (com agrupamento) obtidas da aplicação do método proposto.

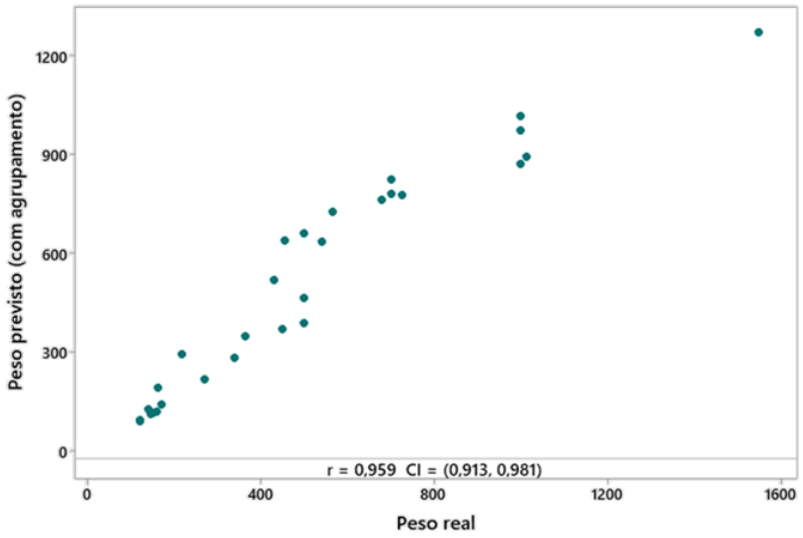


Figura 35 - *Matrix plot* do Peso real e Peso previsto (com agrupamento) (Fonte: autoria própria)

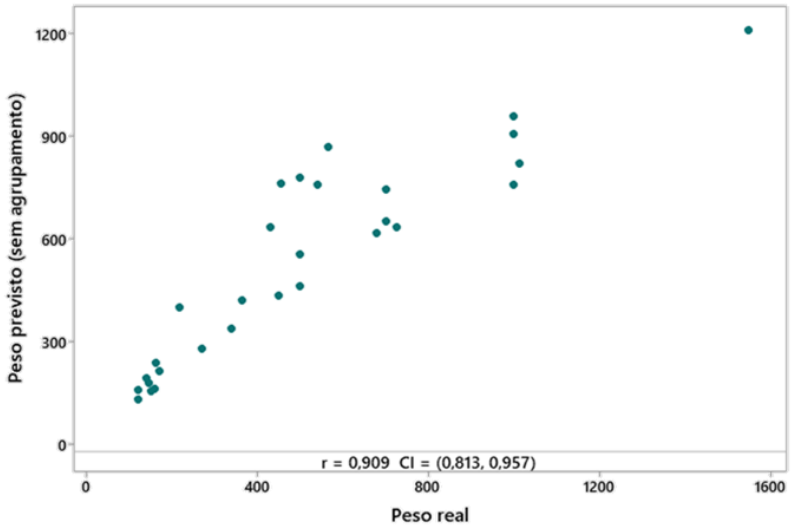


Figura 36 - *Matrix plot* do Peso real e Peso previsto (sem agrupamento) (Fonte: autoria própria)

A Tabela 33 detalha as correlações, e seus respectivos *p-values*, dos valores previstos com e sem agrupamentos com os valores reais do Peso. Assim, conclui-se que, ainda que o conjunto de dados possua um poder preditivo alto originalmente, a inserção da informação do agrupamento é capaz de aumentar o nível de correlação dos resultados, demonstrando maior acurácia na previsão da variável de resposta.

Tabela 33 - Correlação entre os resultados do método proposto e da regressão sem o agrupamento

		Correlação	95% IC para ρ	*p-value*
Peso real	**Peso previsto (com agrupamento)**	0,959	(0,913;0,981)	0,000
	Peso previsto (sem agrupamento)	0,909	(0,813;0,957)	0,000

5. CONCLUSÕES

5.1. Conclusões gerais

O presente estudo buscou elaborar um modelo preditivo do teor de óleos e graxas (TOG) aferido pelo método gravimétrico (TOG-G), tendo como preditores determinadas variáveis do processamento primário de petróleo e as medidas do TOG espectrofotométrico (TOG-S), além de uma nova informação, denominada agrupamento, extraída a partir da própria variável de resposta. Para a extração desta nova variável, realizou-se uma análise combinada da opinião de especialistas e dos quartis dos dados históricos de TOG-G e, então, determinaram-se 4 faixas de TOG-G aos quais os agrupamentos foram associados. Como esta informação de agrupamento é desconhecida para novos dados, fez-se necessário criar um modelo de classificação para que este valor pudesse ser previamente determinado e utilizado como entrada do modelo de previsão o TOG-G. Assim, elaborou-se um método de previsão do TOG-G em duas etapas, onde a primeira envolveu o desenvolvimento deste modelo de classificação para o agrupamento e a segunda utilizou este resultado, juntamente das variáveis de processo e do TOG-S, para construir o modelo de regressão final.

Inicialmente, foram consideradas 81 variáveis de processo e os agrupamentos conhecidos e, a partir de um problema de regressão linear múltipla, este conjunto foi reduzido às 8 variáveis mais significativas que apresentaram *p-value* inferior a 5% (*backward elimination* com $\alpha = 0{,}05$). As variáveis vazão de água produzida para hidrociclone A, tempo de residência no flotador B, total de água produzida Poço 4, produção líquida de óleo Poço 5, e produção líquida de óleo Poço 11 demonstraram contribuir positivamente para o aumento do TOG-G, enquanto que as variáveis total de água produzida Poço 10 e produção líquida de óleo Poço 4 tenderam a reduzir os valores de teor de óleos e graxas. Dentre as variáveis mencionadas, a produção líquida de óleo Poço 4 e produção líquida de óleo Poço 11 foram identificadas como as mais influentes neste cenário. Além disso, a variável TOG-S, que também foi selecionada para o modelo, apresentou contribuição positiva e significativa para a previsão do TOG-G.

É importante destacar que a adequada classificação das novas observações em função dos agrupamentos se mostrou imprescindível para a qualidade de previsão do TOG-G. A utilização conjunta do método de classificação *Random Forest* e a técnica de

balanceamento *undersampling*, produziu resultados satisfatórios para a determinação do valor deste agrupamento.

A partir dos modelos construídos realizou-se a previsão dos valores de TOG-G para novos dados em base diária e base fiscal. Como não é possível realizar uma comparação direta dos resultados da base fiscal com os valores de TOG-G, visto que para este é realizada apenas uma medição por dia, extraíram-se a média e o terceiro quartil dos valores previstos para os 4 horários fiscais por data. Calcularam-se as métricas de MAE, MAPE, R^2 e ρ dos valores diários previstos e da média e do terceiro quartil dos valores fiscais previstos, em relação aos valores reais de TOG-G.

Para comprovar os ganhos com o método proposto, criou-se um novo modelo de regressão linear para a previsão do TOG-G desconsiderando a informação do agrupamento e extraíram-se as mesmas métricas para os resultados obtidos em função dos valores reais de TOG-G. Também foram calculadas as métricas para as medições reais de TOG-S, atualmente utilizada pela plataforma *offshore* como referência de medida em tempo real. A partir destas informações, foi possível verificar que as correlações entre os pares de valores previstos (em base diária ou fiscal) e o TOG-G real superaram o valor de correlação existente tanto entre o TOG-S e o TOG-G, quando entre o TOG-G e os valores previstos a partir do modelo que desconsiderou a informação do agrupamento. Dentre todos, a previsão obtida através do cálculo da média dos valores previstos para a base fiscal apresentou o melhor resultado, com uma correlação de 0,781 com os resultados reais de TOG-G.

A fim de complementar os ganhos com o método proposto, elencou-se um conjunto de dados clássico de regressão linear para previsão do peso de peixes, ou seja, que já apresentasse resultados satisfatórios e altamente correlacionados com os dados reais, e aplicaram-se as etapas do método. Devido à dificuldade em encontrar um problema com a mesma característica do TOG-G, com coletas de dados intermediárias associadas a um resultado único, os passos relativos à transformação dessas informações para dados únicos foram suprimidos. Novamente as métricas de MAE, MAPE, R^2 e ρ foram calculadas, demonstrando que a inclusão do agrupamento foi capaz de aumentar a acurácia do resultado.

Diante disso, constatou-se que a incorporação de informações referentes à resposta de interesse ao conjunto de variáveis preditoras, conforme propõe o método, foi capaz de produzir resultados com correlação estatisticamente significativa em relação ao método gravimétrico, conforme exigência do órgão regulador CONAMA. Além disso, observou-se que pode ser extremamente vantajoso considerar as informações de TOG-G previstos pelo método proposto para tomadas de decisão de ação corretiva ou preventiva do que puramente considerar os resultados do método espectrofotométrico.

5.2. Contribuições do trabalho

Este estudo forneceu informações importantes sobre a contribuição das variáveis do processamento primário de petróleo para os níveis de TOG gravimétrico no descarte da água produzida de uma plataforma FPSO. Além disso, e principalmente, explorou uma estratégia de inclusão de informações da variável de interesse ao modelo de previsão de TOG-G, o que aumentou a capacidade preditiva do modelo e permitiu a criação de resultados estatisticamente correlacionados com o método gravimétrico, homologado pelo CONAMA, para uso local e em tempo real, a partir de informações disponíveis na plataforma. A alta disponibilidade desta informação é essencial para facilitar a tomada de decisão sobre medidas preventivas ou corretivas para controle de danos ambientais e evitar prejuízos financeiros à empresa.

5.3. Sugestões para estudos futuros

Para estudos futuros, sugere-se incorporar ao modelo de previsão de TOG gravimétrico as informações de média e variância associadas aos coeficientes gerados para os modelos das 50 combinações de conjuntos de treinamento e teste, e não somente do da combinação com maior acurácia individual. Além disso, sugere-se a expansão do método para outras plataformas com diferentes características de processo.

REFERÊNCIAS

Agência Brasil. Disponível em: <https://agenciabrasil.ebc.com.br/geral/noticia/2017-11/profissionais-offshore-contam-desafio-de-trabalhar-e-descansar-240-km-da-costa>. Acesso em: 11 nov. 2021.

AKINBINU, V. A. Prediction of fracture gradient from formation pressures and depth using correlation and stepwise multiple regression techniques. **Journal of Petroleum Science and Engineering**, v. 72, n. 1–2, p. 10–17, 2010.

ALLAHYARZADEH-BIDGOLI, A. et al. Energy optimization of an FPSO operating in the Brazilian Pre-salt region. **Energy**, v. 164, p. 390–399, 2018.

APPOLINARIO, F. **Metodologia da ciência: filosofia e prática da pesquisa**. [s.l.] Cencage Learning, 2009.

ARAUJO FILHO, C. F. et al. Monitoramento de teor de óleos e graxas em água descartada no mar usando ciência de dados. v. 2020, n. December 2020, 2020.

ASSOCIATION, A. W. W. **5520 OIL AND GREASE - Standard Methods for Examination of Water and Wastewater**, [s.d.]. Disponível em: <https://www.standardmethods.org/doi/10.2105/SMWW.2882.107>. Acesso em: 26 out. 2021

BALESTRASSI, P. P. et al. Design of experiments on neural network's training for nonlinear time series forecasting. **Neurocomputing**, v. 72, n. 4–6, p. 1160–1178, 2009.

BERTRAND, J. W. M.; FRANSOO, J. C. Operations management research methodologies using quantitative modeling. **International Journal of Operations and Production Management**, v. 22, n. 2, p. 241–264, 2002.

BREIMAN, L. Bagging predictors. **Machine Learning**, v. 24, p. 123–140, 1996.

BREIMAN, L. Random Forests. **Machine Learning 2001 45:1**, v. 45, n. 1, p. 5–32, out. 2001.

BROWNLEE, J. Imbalanced Classification with Python. **Machine Learning Mastery**, p. 463, 2020.

CASTELLANI, M. Competitive co-evolution of multi-layer perceptron classifiers. **Soft Computing**, v. 22, p. 3417–3432, 2018.

CHAU, N. LE; THOAI, T. N.; DAO, T.-P. A hybrid approach of density-based topology, multilayer perceptron, and water cycle-moth flame algorithm for multi-stage optimal design of a

flexure mechanism. **Engineering with Computers**, 2021.

CHEN, H. et al. A combination strategy of random forest and back propagation network for variable selection in spectral calibration. **Chemometrics and Intelligent Laboratory Systems**, v. 182, p. 101–108, 2018.

COMPANY, B. P. **Statistical Review of World EnergyBritish Petroleum Co.** [s.l: s.n.].

CONAMA, C. N. DO M. A. Resolução nº 393, 8 de agosto de 2007. **Ministério do Meio Ambiente**, n. 153, p. 72–73, 2007.

COSTA, T. C. et al. Evaluation of the technical and environmental feasibility of adsorption process to remove water soluble organics from produced water: A review. **Journal of Petroleum Science and Engineering**, v. 208, n. April 2021, 2022.

DAVTYAN, A. et al. Oil production forecast models based on sliding window regression. **Journal of Petroleum Science and Engineering**, v. 195, n. March, p. 107916, 2020.

DE OLIVEIRA, E. C. et al. Uncertainty evaluation in the determination of oil and grease content in produced water by colorimetric method using Monte Carlo Simulation. **Petroleum Science and Technology**, v. 37, n. 4, p. 436–442, 2019.

DOUGHERTY, G. **Pattern Recognition and Classification**. New York: Springer, 2013.

EL-DAHSHAN, E.-S. A.; BASSIOUNI, M. M. Computational intelligence techniques for human brain MRI classification. **International Journal of Imaging Systems and Technology**, v. 28, p. 132–148, 2018.

EVERITT, B.; RENCHER, A. C. **Methods of Multivariate Analysis.** [s.l: s.n.]. v. 45

FAN, J. et al. Confocal microscopy as a new real-time quantification method for oil content in produced water. **Journal of Petroleum Science and Engineering**, v. 167, n. December 2017, p. 54–63, 2018.

GERHARDT, T. E.; SILVEIRA, D. T. **Métodos de pesquisa**. [s.l.] Editora UFRGS, 2019.

HABIBI, M. J. et al. Prediction of permeability in dual fracture media by multivariate regression analysis. **Journal of Petroleum Science and Engineering**, v. 120, p. 194–201, 2014.

HAN, Y.; ZHEN, X.; HUANG, Y. Drift-off warning limits for dynamically positioned FPSO and Deepwater Artificial Seabed (DAS) coupling system. **Ocean Engineering**, v. 237, n. August, p. 109662, 2021.

HASSANAT, A. B. et al. Solving the Problem of the K Parameter in the KNN Classifier Using an Ensemble Learning Approach. **International Journal of Computer Science and Information Security,** v. 12, 2014.

HASTIE, T. ET. ALL. Springer Series in Statistics The Elements of Statistical Learning. **The Mathematical Intelligencer**, v. 27, n. 2, p. 83–85, 2009.

HASTIE, T.; TIBSHIRANI, R.; FRIEDMAN, J. **The Elements of Statistical Learning - Data Mining, Inference, and Prediction**. Second Edi ed. [s.l.] Springer, 2017.

HAYKIN, S. **Neural Networks and Learning Machines**. [s.l.] Prentice Hall, 2009.

HEGDE, C.; MILLWATER, H.; GRAY, K. Classification of drilling stick slip severity using machine learning. **Journal of Petroleum Science and Engineering**, v. 179, n. May, p. 1023–1036, 2019.

HOSMER JR., D. W.; LEMESHOW, S.; STURDIVANT, R. X. **Applied Logistic Regression**. Third Edit ed. Hoboken, New Jersey: John Wiley & Sons, Inc., 2013.

JOHNSON, R. A.; WICHERN, D. W. **Applied multivariate statistical analysis**. 6th. ed. [s.l.] Prentice Hall, 2007.

JÚNIOR, J. F. H. et al. **Multivariate Data Analysis**. 8. ed. [s.l.] Cengage Learning, 2019.

KLEMZ, A. C. et al. Oilfield produced water treatment by liquid-liquid extraction: A review. **Journal of Petroleum Science and Engineering**, v. 199, n. July 2020, 2021.

LAZRI, M.; AMEUR, S. Combination of support vector machine, artificial neural network and random forest for improving the classification of convective and stratiform rain using spectral features of SEVIRI data. **Atmospheric Research**, v. 203, p. 118–129, 2018.

LEE, K.; NEFF, J. **Produced Water**. 1. ed. [s.l.] Springer, 2011.

LEE, T. H.; ULLAH, A.; WANG, R. Bootstrap Aggregating and Random Forest. **Advanced Studies in Theoretical and Applied Econometrics**, v. 52, p. 389–429, 2020.

LIN, S.-K. et al. Classification of patients with Alzheimer's disease using the arterial pulse spectrum and a multilayer-perceptron analysis. **Scientifica Reports**, v. 11, 2021.

LIN, W. C. et al. Clustering-based undersampling in class-imbalanced data. **Information Sciences**, v. 409–410, p. 17–26, 2017.

LV, P. et al. Optimization of chemical agents for removing dispersed oil from produced water in Z oilfield. **Petroleum Science and Technology**, v. 35, n. 12, p. 1285–1289, 2017.

MARDIA, K. V.; KENT, J. T.; BIBBY, J. M. **Multivariate Analysis**. [s.l.] Academic Press, 1995.

MARINS, M. A. et al. Fault detection and classification in oil wells and production/service lines using random forest. **Journal of Petroleum Science and Engineering**, v. 197, n. March 2020, p. 107879, 2021.

MIGUEL, P. A. C. et al. **Metodologia de Pesquisa em Engenharia de Produção e Gestão de Operações**Elsevier, , 2014.

MÜLLER, A. C.; GUIDO, S. **Introduction to Machine Learning with Python a guide for datat scientists**. 1st. ed. Gravenstein Highway North, Sebastopol: O'Reilly Media, Inc., 2016.

MÜLLER, A. C.; GUIDO, S. **Introduction to Machine Learning with Python**. [s.l: s.n.].

ROKACH, L. **Pattern Classification using ensemble methods**. [s.l.] World Scientific, 2010.

S.A., P. **PE-3UBC-03001 - Teor de Óleo e Graxa em Água por Gravimetria**, 2019a.

S.A., P. **PE-3UBC-02899 - Teor de Óleo e Graxa em Água por Espectrofotometria de Absorção Molecular**, 2019b.

S.A., P. PE-3UBC-02967 - Teor de Óleo e Graxa em Água por Infravermelho com Eracheck ECO. 2020.

SANQUETTA, C. R. et al. Volume estimation of Cryptomeria japonica logs in southern Brazil using artificial intelligence models. **Southern Forests**, v. 80, n. 1, p. 29–36, 2018.

SHARMA, S. **Applied Multivariate Techniques**. [s.l.] John Wiley & Sons, Inc, 1996.

SILVA, I. N. DA et al. **Artificial Neural Networks: A Practical Course**. [s.l.] Springer, 2016.

SIMÕES, L. D. et al. A power transformer differential protection based onsupport vector machine and wavelet transform. **Electric Power Systems Research**, v. 197, 2021.

SLEITI, A. K. et al. Comprehensive assessment and evaluation of correlations for gas-oil ratio, oil formation volume factor, gas viscosity, and gas density utilized in gas kick detection. **Journal of Petroleum Science and Engineering**, v. 207, n. May, p. 109135, 2021.

TRIGGIA, A. A. et al. **Fundamentos de Engenharia de Petróleo**. Rio de Janeiro: Interciência, Editora, 2001.

UDDIN, S. et al. Comparative performance analysis of K-nearest neighbour (KNN) algorithm and its different variants for disease prediction. **Scientifics Reports**, v. 12, 2022.

WANG, H.; MOAYEDI, H.; FOONG, L. K. Genetic algorithm hybridized with multilayer perceptron to have an economical slope stability design. **Engineering with Computers**, v. 37, p. 3067–3078, 2020.

YAP, B. W. et al. An application of oversampling, undersampling, bagging and boosting in handling imbalanced datasets. **Lecture Notes in Electrical Engineering**, v. 285 LNEE, p. 13–22, 2014.

YEGNANARAYANA, B. **Artificial Neural Networks**. 11. ed. [s.l.] New Delhi: Prentice-Hall of India Private Limited, 2005.

ZHANG, C.; MA, Y. **Ensemble Machine Learning**. [s.l.] Springer, 2012.

ZHANG, S. Nearest neighbor selection for iteratively kNN imputation. **Journal of Systems and Software**, v. 85, n. 11, p. 2541–2552, 2012.

ZHANG, S. et al. A novel kNN algorithm with data-driven k parameter computation. **Pattern Recognition Letters**, v. 109, p. 44–54, 2018a.

ZHANG, S. et al. Efficient kNN classification with different numbers of nearest neighbors. **IEEE Transactions on Neural Networks and Learning Systems**, v. 29, p. 1774–1785, 2018b.

ZUO, W.; ZHANG, D.; WANG, K. On kernel difference-weighted k-nearest neighbor classification. **Pattern Analysis and Applications**, v. 11, p. 247–257, 2008.

APÊNDICE A – Descrição das variáveis da base de dados original

Código da variável	Descrição	Controlável
x_1	Método de elevação Poço1	Não
x_2	Método de elevação Poço 2	Não
x_3	Método de elevação Poço 3	Não
x_4	Método de elevação Poço 4	Não
x_5	Método de elevação Poço 5	Não
x_6	Método de elevação Poço 6	Não
x_7	Método de elevação Poço 7	Não
x_8	Método de elevação Poço 8	Não
x_9	Método de elevação Poço 9	Não
x_{10}	Método de elevação Poço 10	Não
x_{11}	Método de elevação Poço 11	Não
x_{12}	Método de elevação Poço 12	Não
x_{13}	Método de elevação Poço 13	Não
x_{14}	Pressão de chegada do Poço 7	Sim
x_{15}	Pressão de chegada do Poço 9	Sim
x_{16}	Pressão de chegada do Poço 5	Sim
x_{17}	Pressão de chegada do Poço 3	Sim
x_{18}	Pressão de chegada do Poço 6	Sim
x_{19}	Pressão de chegada do Poço 2	Sim
x_{20}	Pressão de chegada do Poço 4	Sim
x_{21}	Pressão de chegada do Poço 8	Sim
x_{22}	Pressão de chegada duto produção 1	Sim
x_{23}	Pressão de chegada duto teste 1 produção	Sim
x_{24}	Pressão de chegada duto produção 2	Sim
x_{25}	Pressão de chegada duto teste produção 2	Sim
x_{26}	Temperatura de chegada do Poço 7	Não
x_{27}	Temperatura de chegada do Poço 9	Não
x_{28}	Temperatura de chegada do Poço 5	Não
x_{29}	Temperatura de chegada do Poço 3	Não
x_{30}	Temperatura de chegada do Poço 6	Não
x_{31}	Temperatura de chegada do Poço 2	Não
x_{32}	Temperatura de chegada do Poço 4	Não
x_{33}	Temperatura de chegada do Poço 8	Não
x_{34}	Temperatura de chegada duto teste produção 1	Não
x_{35}	Temperatura de chegada duto teste produção 2	Não
x_{36}	Vazão de óleo tramo 1	Sim
x_{37}	Vazão de óleo tramo 2	Sim
x_{38}	Vazão de óleo tramo 3	Sim
x_{39}	Vazão de gás separador LP	Sim
x_{40}	Vazão de gás Separador HP	Sim
x_{41}	Separador HP - Vazão de óleo tramo A	Sim

Código da variável	Descrição	Controlável
x_{42}	Separador HP - Vazão de óleo tramo B	Sim
x_{43}	Vazão de água separador de teste tramo A	Sim
x_{44}	Vazão de água separador de teste tramo B	Sim
x_{45}	Vazão de óleo separador de teste tramo A	Sim
x_{46}	Vazão de óleo separador de teste tramo B	Sim
x_{47}	Vazão de gás separador de teste	Sim
x_{48}	Volume corrigido de gás no separador de teste na hora anterior	Sim
x_{49}	Corrente Trafo B do Pré-TO	Sim
x_{50}	Corrente Trafo C do Pré-TO	Sim
x_{51}	Corrente Trafo A TO	Sim
x_{52}	Corrente Trafo B TO	Sim
x_{53}	Corrente Trafo C TO	Sim
x_{54}	Nível óleo settling B	Sim
x_{55}	Nível separador de produção HP	Sim
x_{56}	Nível interface Pré-TO	Sim
x_{57}	Nível separador de produção LP	Sim
x_{58}	Nível interface TO	Sim
x_{59}	Nível gás/óleo separador de teste	Sim
x_{60}	Nível água/óleo separador de teste	Sim
x_{61}	BSW a 11m do fundo do *settling* A	Não
x_{62}	BSW a 11m do fundo do *settling* B	Não
x_{63}	BSW 12m do fundo do *settling* A	Não
x_{64}	BSW 12m do fundo do *settling* B	Não
x_{65}	BSW 10m do fundo do *settling* A	Não
x_{66}	BSW 10m do fundo do *settling* B	Não
x_{67}	BSW 9m do fundo do *settling* A	Não
x_{68}	BSW 9m do fundo do *settling* B	Não
x_{69}	BSW 13m do fundo do *settling* A	Não
x_{70}	Volume tanque de *slop* boreste	Sim
x_{71}	Nível tanque de *slop* boreste	Sim
x_{72}	Nível tanque de *slop* boreste	Sim
x_{73}	Pressão separador de produção LP	Sim
x_{74}	Pressão TO	Sim
x_{75}	Separador HP - Pressão medidor de gás	Sim
x_{76}	Pressão separador de produção HP	Sim
x_{77}	Pressão Pré-TO	Sim
x_{78}	Pressão saída de gás separador de teste	Sim
x_{79}	Pressão saída de gás separador de teste	Sim
x_{80}	Pressão saída de gás *settling* A	Sim
x_{81}	Pressão saída de gás *settling* B	Sim
x_{82}	Pressão descarga bomba A de água *settling* A	Sim
x_{83}	Pressão descarga bomba C de água *settling* B	Sim
x_{84}	Pressão descarga bomba B de água *settling* A	Sim
x_{85}	Pressão descarga bomba D de água *settling* B	Sim
x_{86}	Pressão saída de óleo TO	Sim
x_{87}	Pressão saída de gás separador LP	Sim

Código da variável	Descrição	Controlável
x_{88}	Pressão do tanque de *slop* boreste	Sim
x_{89}	Temperatura entrada separador de produção HP	Sim
x_{90}	Temperatura TO	Sim
x_{91}	Temperatura saída de óleo bruto	Sim
x_{92}	Temperatura saída de óleo	Sim
x_{93}	Temperatura saída de gás separador de teste	Sim
x_{94}	Temperatura entrada separador de produção LP	Sim
x_{95}	Temperatura do tanque de *slop* boreste	Sim
x_{96}	Temperatura do *settling* tanque 4 central	Sim
x_{97}	Temperatura do *settling* tanque 3 central	Sim
x_{98}	Pressão descarga bomba *cleaning* do *slop*	Sim
x_{99}	TOG no descarte da água produzida	Não
x_{100}	Vazão de descarte da água produzida	Sim
x_{101}	Vazão de água produzida para hidrociclone A	Sim
x_{102}	Vazão de água produzida entrada hidrociclone B	Sim
x_{103}	Nível na câmara de óleo do flotador A	Sim
x_{104}	Nível na câmara de óleo do flotador B	Sim
x_{105}	Nível da câmara de água flotador A	Sim
x_{106}	Nível da câmara de água flotador B	Sim
x_{107}	Tempo de residência flotador A	Sim
x_{108}	Tempo de residência flotador B	Sim
x_{109}	Pressão diferencial entrada/saída de água hidrociclone A	Sim
x_{110}	Pressão diferencial entrada/saída de água hidrociclone B	Sim
x_{111}	Pressão flotador A	Sim
x_{112}	Pressão flotador B	Sim
x_{113}	Pressão flotador A	Sim
x_{114}	Pressão flotador B	Sim
x_{115}	Desemulsificante *Topside*	Sim
x_{116}	Desemulsificante *Subsea*	Sim
x_{117}	Inibidor de incrustação *Topside*	Sim
x_{118}	Polieletrólito	Sim
x_{119}	Total Água Produzida Poço 1	Sim
x_{120}	Total Água Produzida Poço 2	Sim
x_{121}	Total Água Produzida Poço 3	Sim
x_{122}	Total Água Produzida Poço 4	Sim
x_{123}	Total Água Produzida Poço 5	Sim
x_{124}	Total Água Produzida Poço 6	Sim
x_{125}	Total Água Produzida Poço 7	Sim
x_{126}	Total Água Produzida Poço 8	Sim
x_{127}	Total Água Produzida Poço 9	Sim
x_{128}	Total Água Produzida Poço 10	Sim
x_{129}	Total Água Produzida Poço 11	Sim
x_{130}	Total Água Produzida Poço 12	Sim
x_{131}	Total Água Produzida Poço 13	Sim
x_{132}	BSW Poço 1	Não
x_{133}	BSW Poço 2	Não

Código da variável	Descrição	Controlável
x_{134}	BSW Poço 3	Não
x_{135}	BSW Poço 4	Não
x_{136}	BSW Poço 5	Não
x_{137}	BSW Poço 6	Não
x_{138}	BSW Poço 7	Não
x_{139}	BSW Poço 8	Não
x_{140}	BSW Poço 9	Não
x_{141}	BSW Poço 10	Não
x_{142}	BSW Poço 11	Não
x_{143}	BSW Poço 12	Não
x_{144}	BSW Poço 13	Não
x_{145}	Produção Bruta Poço 1	Sim
x_{146}	Produção Bruta Poço 2	Sim
x_{147}	Produção Bruta Poço 3	Sim
x_{148}	Produção Bruta Poço 4	Sim
x_{149}	Produção Bruta Poço 5	Sim
x_{150}	Produção Bruta Poço 6	Sim
x_{151}	Produção Bruta Poço 7	Sim
x_{152}	Produção Bruta Poço 8	Sim
x_{153}	Produção Bruta Poço 9	Sim
x_{154}	Produção Bruta Poço 10	Sim
x_{155}	Produção Bruta Poço 11	Sim
x_{156}	Produção Bruta Poço 12	Sim
x_{157}	Produção Bruta Poço 13	Sim
x_{158}	Produção Líquida de Óleo Poço 1	Sim
x_{159}	Produção Líquida de Óleo Poço 2	Sim
x_{160}	Produção Líquida de Óleo Poço 3	Sim
x_{161}	Produção Líquida de Óleo Poço 4	Sim
x_{162}	Produção Líquida de Óleo Poço 5	Sim
x_{163}	Produção Líquida de Óleo Poço 6	Sim
x_{164}	Produção Líquida de Óleo Poço 7	Sim
x_{165}	Produção Líquida de Óleo Poço 8	Sim
x_{166}	Produção Líquida de Óleo Poço 9	Sim
x_{167}	Produção Líquida de Óleo Poço 10	Sim
x_{168}	Produção Líquida de Óleo Poço 11	Sim
x_{169}	Produção Líquida de Óleo Poço 12	Sim
x_{170}	Produção Líquida de Óleo Poço 13	Sim
x_{171}	Total Água Descartada	Sim
x_{172}	Total Água Produzida	Sim
x_{173}	Total Produção Bruta	Sim
x_{174}	Total Produção Líquida de Óleo	Sim
x_{175}	TOG *eracheck*	Não
x_{176}	TOG espectrofotométrico	Não
x_{177}	TOG gravimétrico	Não

APÊNDICE B – Correlação de Pearson entre as variáveis preditoras da base de dados original e a variável TOG-G

Código da variável	Correlação	Intervalo de confiança (95%) para ρ	p-value
x_1	0,070	(0,010;0,130)	0,023
x_2	-0,077	(-0,136;-0,016)	0,013
x_3	0,141	(0,081;0,200)	0,000
x_4	-0,004	(-0,064;0,057)	0,904
x_5	0,098	(0,038;0,158)	0,001
x_6	0,217	(0,159;0,274)	0,000
x_7	-0,009	(-0,069;0,052)	0,782
x_8	-0,168	(-0,226;-0,108)	0,000
x_9	0,080	(0,020;0,140)	0,009
x_{10}	0,001	(-0,059;0,062)	0,965
x_{11}	0,089	(0,028;0,148)	0,004
x_{12}	0,065	(0,004;0,125)	0,035
x_{13}	0,025	(-0,036;0,085)	0,420
x_{14}	0,463	(0,414;0,509)	0,000
x_{15}	0,329	(0,274;0382)	0,000
x_{16}	0,231	(0,173;0,288)	0,000
x_{17}	0,280	(0,223;0,334)	0,000
x_{18}	0,052	(-0,008;0,112)	0,092
x_{19}	0,201	(0,142;0,258)	0,000
x_{20}	-0,237	(-0,294;-0,179)	0,000
x_{21}	0,133	(0,073;0,192)	0,000
x_{22}	0,063	(0,003;0,123)	0,041
x_{23}	0,090	(0,030;0,150)	0,003
x_{24}	0,321	(0,265;0,374)	0,000
x_{25}	0,415	(0,364;0,464)	0,000
x_{26}	0,239	(0,181;0,295)	0,000
x_{27}	0,492	(0,445;0,537)	0,000
x_{28}	0,145	(0,085;0,204)	0,000
x_{29}	0,071	(0,011;0,131)	0,021
x_{30}	0,211	(0,153;0,268)	0,000
x_{31}	0,053	(-0,007;0,114)	0,084
x_{32}	0,022	(-0,039;0,082)	0,479
x_{33}	0,211	(0,153;0,268)	0,000
x_{34}	-0,198	(-0,256;-0,140)	0,000
x_{35}	-0,100	(-0,159;-0,039)	0,001

Código da variável	Correlação	Intervalo de confiança (95%) para ρ	p-value
x_{36}	0,180	(0,120;0,238)	0,000
x_{37}	0,289	(0,232;0,343)	0,000
x_{38}	0,140	(0,080;0,199)	0,000
x_{39}	0,512	(0,466;0,555)	0,000
x_{40}	0,322	(0,267;0,376)	0,000
x_{41}	0,420	(0,368;0,468)	0,000
x_{42}	0,226	(0,168;0,283)	0,000
x_{43}	0,033	(-0,028;0,093)	0,292
x_{44}	0,009	(-0,051;0,070)	0,761
x_{45}	-0,127	(-0,186;-0,067)	0,000
x_{46}	0,302	(0,246;0,356)	0,000
x_{47}	0,080	(0,020;0,140)	0,009
x_{48}	0,085	(0,024;0,144)	0,006
x_{49}	0,148	(0,088;0,207)	0,000
x_{50}	0,256	(0,198;0,312)	0,000
x_{51}	0,122	(0,058;0,185)	0,000
x_{52}	0,283	(0,227;0,338)	0,000
x_{53}	0,102	(0,041;0,161)	0,001
x_{54}	-0,325	(-0,385;-0,263)	0,000
x_{55}	-0,461	(-0,508;-0,412)	0,000
x_{56}	0,416	(0,363;0,466)	0,000
x_{57}	-0,410	(-0,459;-0,358)	0,000
x_{58}	0,190	(0,127;0,251)	0,000
x_{59}	0,081	(0,021;0,141)	0,009
x_{60}	-0,074	(-0,134;-0,014)	0,016
x_{61}	0,122	(0,043;0,199)	0,002
x_{62}	0,124	(-0,005;0,249)	0,059
x_{63}	0,132	(0,029;0,232)	0,012
x_{64}	-0,023	(-0,280;0,236)	0,862
x_{65}	-0,340	(-0,395;-0,282)	0,000
x_{66}	0,146	(0,019;0,269)	0,025
x_{67}	-0,067	(-0,127;-0,007)	0,029
x_{68}	0,338	(0,264;0,408)	0,000
x_{69}	0,148	(-0,013;0,301)	0,071
x_{70}	-0,074	(-0,134;-0,014)	0,016
x_{71}	-0,083	(-0,143;-0,023)	0,007
x_{72}	-0,083	(-0,143;-0,023)	0,007
x_{73}	0,420	(0,367;0,469)	0,000
x_{74}	0,323	(0,268;0,376)	0,000

Código da variável	Correlação	Intervalo de confiança (95%) para ρ	p-value
x_{75}	0,184	(0,125;0,242)	0,000
x_{76}	0,344	(0,290;0,396)	0,000
x_{77}	0,298	(0,240;0,354)	0,000
x_{78}	0,268	(0,211;0,324)	0,000
x_{79}	0,270	(0,213;0,326	0,000
x_{80}	0,231	(0,173;0,288)	0,000
x_{81}	0,044	(-0,026;0,114)	0,219
x_{82}	-0,616	(-0,669;-0,556)	0,000
x_{83}	0,474	(0,338;0,590)	0,000
x_{84}	0,091	(0,013;0,168)	0,023
x_{85}	0,052	(-0,043;0,147)	0,284
x_{86}	0,297	(0,240;0,351)	0,000
x_{87}	-0,035	(-0,095;0,026)	0,259
x_{88}	0,166	(0,107;0,224)	0,000
x_{89}	-0,107	(-0,167;-0,047)	0,001
x_{90}	0,304	(0,248;0,358)	0,000
x_{91}	0,174	(0,114;0,232)	0,000
x_{92}	0,072	(0,012;0,132)	0,019
x_{93}	0,141	(0,081;0,200)	0,000
x_{94}	-0,184	(-0,241;-0,124)	0,000
x_{95}	0,183	(0,124;0,241)	0,000
x_{96}	0,359	(0,305;0,410)	0,000
x_{97}	-0,386	(-0,437;-0,334)	0,000
x_{98}	0,091	(0,031;0,151)	0,003
x_{99}	0,395	(0,343;0,445)	0,000
x_{100}	-0,115	(-0,174;-0,055)	0,000
x_{101}	-0,008	(-0,068;0,053)	0,806
x_{102}	0,053	(-0,010;0,116)	0,101
x_{103}	-0,184	(-0,242;-0,124)	0,000
x_{104}	0,163	(0,103;0,221)	0,000
x_{105}	-0,002	(-0,063;0,058)	0,944
x_{106}	-0,166	(-0,255;-0,107)	0,000
x_{107}	-0,110	(-0,170;-0,050)	0,000
x_{108}	-0,193	(-0,251;-0,134)	0,000
x_{109}	0,043	(-0,020;0,105)	0,181
x_{110}	-0,083	(-0,144;-0,022)	0,008
x_{111}	0,034	(-0,027;0,095)	0,270
x_{112}	0,226	(0,168;0,282)	0,000
x_{113}	0,029	(-0,032;0,090)	0,353

Código da variável	Correlação	Intervalo de confiança (95%) para ρ	p-value
x_{114}	0,016	(-0,044;0,077)	0,599
x_{115}	-0,098	(-0,164;-0,031)	0,004
x_{116}	0,004	(-0,056;0,065)	0,888
x_{117}	-0,115	(-0,183;-0,045)	0,001
x_{118}	0,190	(0,131;0,248)	0,000
x_{119}	-0,454	(-0,501;-0,405)	0,000
x_{120}	-0,570	(-0,609;-0,528)	0,000
x_{121}	0,036	(-0,024;0,097)	0,241
x_{122}	-0,074	(-0,134;-0,013)	0,017
x_{123}	0,129	(0,069;0,188)	0,000
x_{124}	0,229	(0,170;0,285)	0,000
x_{125}	-0,283	(-0,338;-0,226)	0,000
x_{126}	-0,370	(-0,422;-0,317)	0,000
x_{127}	0,210	(0,151;0,267)	0,000
x_{128}	-0,509	(-0,552;-0,462)	0,000
x_{129}	0,126	(0,066;0185)	0,000
x_{130}	-0,494	(-0,539;-0,447)	0,000
x_{131}	-0,410	(-0,459;-0,358)	0,000
x_{132}	-0,524	(-0,566;-0,479)	0,000
x_{133}	-0,554	(-0,594;-0,510)	0,000
x_{134}	-0,391	(-0,441;-0,338)	0,000
x_{135}	-0,271	(-0,326;-0,214)	0,000
x_{136}	-0,107	(-0,166;-0,047)	0,001
x_{137}	-0,263	(-0,318;-0,205)	0,000
x_{138}	-0,360	(-0,412;-0,306)	0,000
x_{139}	-0,565	(-0,605;-0,522)	0,000
x_{140}	-0,469	(-0,515;-0,420)	0,000
x_{141}	-0,507	(-0,550;-0,460)	0,000
x_{142}	0,037	(-0,024;0,097)	0,236
x_{143}	-0,541	(-0,582;-0,496)	0,000
x_{144}	-0,540	(-0,582;-0,496)	0,000
x_{145}	0,461	(0,412;0,507)	0,000
x_{146}	0,295	(0,239;0,349)	0,000
x_{147}	0,076	(0,016;0,136)	0,013
x_{148}	0,174	(0,115;0,232)	0,000
x_{149}	0,147	(0,087;0,206)	0,000
x_{150}	0,240	(0,182;0,297)	0,000
x_{151}	-0,040	(-0,100;0,020)	0,194
x_{152}	0,062	(0,001;0,122)	0,046

Código da variável	Correlação	Intervalo de confiança (95%) para ρ	p-value
x_{153}	0,390	(0,337;0,440)	0,000
x_{154}	0,034	(-0,026;0,095)	0,265
x_{155}	0,225	(0,167;0,282)	0,000
x_{156}	-0,339	(-0,391;-0,284)	0,000
x_{157}	0,298	(0,242;0,352)	0,000
x_{158}	0,489	(0,441;0,533)	0,000
x_{159}	0,374	(0,320;0,425)	0,000
x_{160}	0,142	(0,082;0,201)	0,000
x_{161}	0,184	(0,125;0,242)	0,000
x_{162}	0,110	(0,050;0,170)	0,000
x_{163}	0,284	(0,228;0,339)	0,000
x_{164}	0,267	(0,210;0,322)	0,000
x_{165}	0,163	(0,104;0,221)	0,000
x_{166}	0,438	(0,388;0,485)	0,000
x_{167}	0,363	(0,309;0,415)	0,000
x_{168}	0,150	(0,090;0,209)	0,000
x_{169}	0,163	(0,104;0,222)	0,000
x_{170}	0,431	(0,381;0,479)	0,000
x_{171}	-0,111	(-0,170;-0,051)	0,000
x_{172}	-0,118	(-0,177;-0,058)	0,000
x_{173}	0,0446	(0,396;0,493)	0,000
x_{174}	0,445	(0,395;0,492)	0,000
x_{175}	0,039	(-0,063;0,140)	0,454
x_{176}	0,571	(0,529;0,610)	0,000