VALIDAÇÃO DE TESTES DE PRODUÇÃO DE POÇOS DE PETRÓLEO BASEADA EM MINERAÇÃO DE DADOS

Maria Clara Machado de Almeida Duque

Dissertação de Mestrado apresentada ao Programa de Pós-graduação em Engenharia de Produção, COPPE, da Universidade Federal do Rio de Janeiro, como parte dos requisitos necessários à obtenção do título de Mestre em Engenharia de Produção.

Orientador: Virgílio José Martins Ferreira Filho

Rio de Janeiro
Março de 2019

VALIDAÇÃO DE TESTES DE PRODUÇÃO DE POÇOS DE PETRÓLEO BASEADA EM MINERAÇÃO DE DADOS

Maria Clara Machado de Almeida Duque

DISSERTAÇÃO SUBMETIDA AO CORPO DOCENTE DO INSTITUTO ALBERTO LUIZ COIMBRA DE PÓS-GRADUAÇÃO E PESQUISA DE ENGENHARIA (COPPE) DA UNIVERSIDADE FEDERAL DO RIO DE JANEIRO COMO PARTE DOS REQUISITOS NECESSÁRIOS PARA A OBTENÇÃO DO GRAU DE MESTRE EM CIÊNCIAS EM ENGENHARIA DE PRODUÇÃO.

Examinada por:

_____

Prof. Virgílio José Martins Ferreira Filho, D.Sc.

_____

Prof[a]. Juliana Souza Baioco, D.Sc.

_____

Prof. Alexandre Gonçalves Evsukoff, D.Sc.

RIO DE JANEIRO, RJ - BRASIL

MARÇO DE 2019

Duque, Maria Clara Machado de Almeida

Validação de Testes de Produção de Poços de Petróleo baseada em Mineração de Dados/ Maria Clara Machado de Almeida Duque. – Rio de Janeiro: UFRJ/COPPE, 2019.

XIV, 150 p.: il.; 29,7 cm.

Orientador: Virgílio José Martins Ferreira Filho

Dissertação (mestrado) – UFRJ/ COPPE/ Programa de Engenharia de Produção, 2019.

Referências Bibliográficas: p. 97-100.

1. Validação de Testes. 2. Mineração de Dados. 3. Produção de Petróleo. I. Ferreira Filho, Virgílio José Martins. II. Universidade Federal do Rio de Janeiro, COPPE, Programa de Engenharia de Produção. III. Título.

AGRADECIMENTOS

A Deus, por toda força me dada para enfrentar os momentos difíceis e por todas as oportunidades oferecidas.

A minha família, em especial aos meus pais, por todo incentivo e por estarem sempre presentes me apoiando.

Ao professor Virgílio pela orientação, oportunidades e ensinamentos. Estes anos no laboratório me fizeram crescer muito no aspecto acadêmico e profissional. A professora Juliana também por tudo que me ajudou durante o período do projeto.

Aos professores Juliana e Alexandre pela disponibilidade de tempo, solicitude ao terem aceitado fazer parte da banca avaliadora.

Aos meus amigos por todo o apoio, em especial aos amigos do laboratório pelo incentivo e paciência neste período difícil.

Aos funcionários do LORDE/SAGE, em especial ao seu Zé e a Soyla pela disponibilidade, atenção e carinho.

Aos professores do Programa de Engenharia de Produção – PEP/COPPE, por todo conhecimento proporcionado ao longo desses dois anos de mestrado.

À CAPES pelo apoio financeiro.

Resumo da Dissertação apresentada à COPPE/UFRJ como parte dos requisitos necessários para a obtenção do grau de Mestre em Ciências (M.Sc.)

VALIDAÇÃO DE TESTES DE PRODUÇÃO DE POÇOS DE PETRÓLEO BASEADA EM MINERAÇÃO DE DADOS

Maria Clara Machado de Almeida Duque

Março/2019

Orientador: Virgílio José Martins Ferreira Filho
Programa: Engenharia de Produção

Durante a produção de campos de petróleo, testes de produção são conduzidos periodicamente em poços de petróleo para identificar as condições correntes de produção de cada poço. Após a finalização dos testes, estes são avaliados pela equipe responsável e, de acordo com as informações obtidas, podem ser validados ou não. O objetivo desse trabalho é criar ferramentas de validação de teste de produção, baseadas em mineração de dados para auxiliar no processo de validação em tempo real. A metodologia proposta é dividida em três etapas principais. Na primeira, um pré-processamento é feito para identificação de dados anômalos, utilizando os métodos LOF (*Local Outlier Factor*), Z-score modificado e da média das distâncias. Após isso, na segunda etapa, modelos preditivos de classificação são analisados para caracterizar um teste de produção como válido e inválido, de acordo com informações do histórico de produção do poço. Na terceira etapa, são aplicados modelos de regressão para previsão das variáveis de vazão de óleo, água e gás. Nesta parte ainda, um intervalo de predição para cada variável é construído através da técnica de amostragem *bootstrap*. A metodologia proposta foi aplicada em 13 poços representativos de um campo de petróleo brasileiro. As técnicas desenvolvidas facilitam o processo de tomada de decisão nas atividades de produção de petróleo.

Abstract of Dissertation presented to COPPE/UFRJ as a partial fulfillment of the requirements for the degree of Master of Science (M.Sc.)

WELL PRODUCTION TEST VALIDATION BASED ON DATA MINING

Maria Clara Machado de Almeida Duque

March/2019

Advisor: Virgílio José Martins Ferreira Filho

Department: Industrial Engineering

During production of petroleum fields, production tests are frequently conducted in each well to identify the current conditions of well production. After tests are done, they are evaluated by the responsible team and, according to the information obtained, can be validated or not. The objective of this work is to create production validation tools based on data mining to assist real-time validation process. The proposed methodology is divided into three main stages. In the first one, a preprocessing is done to identify anomalous data, using the Local Outlier Factor (LOF), modified Z-score and average distances. After that, in the second step, predictive models of classification are analyzed to characterize production test as valid and not valid, according to information of the production history of the well. In third step, regression models are applied to predict the oil, water and gas flow variables. In this part, a prediction interval for each variable is constructed using the bootstrap sampling technique. The proposed methodology was applied in 13 representative wells of a Brazilian oil field. The developed techniques collaborate with the decision-making process in oil production activities.

Sumário

LISTA DE FIGURAS

LISTA DE TABELAS

LISTA DE ABREVIATURAS E SIGLAS

ANP - Agência Nacional de Petróleo, Gás Natural e Biocombustíveis

AUC - *Area Under the ROC Curve*

BSW – B*asic Sediment Water*

IPR - *Inflow Performance Relationship*

KNN - K-Vizinhos mais próximos

MLR – Regressão Linear Múltipla

NB - Classificador *Naive Bayes*

PDG - *Pressure Downhole Gauge*

RF - Floresta Aleatória

RL - Regressão Logística

RFR – Floresta Aleatória para Regressão

RGLI – Razão gás-líquido

RGO – Razão gás-óleo

RT – Árvore de Regressão

SVM - Máquina de Vetores de Suporte

SVR – Regressão por Vetores de Suporte

TPT – *Tree Pressure and Temperature*

1 INTRODUÇÃO

1.1 Contexto e Motivação

O processo de produção de campos de petróleo é uma tarefa complexa, especialmente em um cenário atual com o aumento da produção no ambiente *offshore* em águas cada vez mais profundas. De forma geral, campos em produção utilizam vários poços que são responsáveis por retirar óleo e/ou gás do reservatório e conduzi-lo até um sistema de coleta dessa produção, através de diferentes arranjos e equipamentos que permitem o transporte de fluido.

A produção de poços é uma atividade dinâmica. O fluido produzido é normalmente uma mistura multifásica complexa composta por óleo, água e gás. Além disso, padrões e características inerentes ao poço, tais como pressões e vazões, estão sempre modificando, de forma que é necessário avaliar constantemente o comportamento do poço.

Entretanto, existem dificuldades enfrentadas na avaliação do comportamento dos poços. Em um sistema de produção de petróleo, os fluidos produzidos pelos poços seguem para um coletor comum na plataforma e somente depois passam por um processo de separação de fases, em que se tem como resultado as correntes de água, óleo e gás separadas. Assim, é possível se ter medidas das vazões das correntes para um conjunto de poços que passaram pela mesma unidade de separação, mas não existe um controle individual preciso da produção de cada poço. É somente quando testes de produção são conduzidos que se há uma visão real de como cada poço está produzindo.

Testes de produção são realizados periodicamente em cada poço com o objetivo de identificar condições correntes de produção e possíveis problemas, além de analisar possíveis oportunidades para aumentar a produção. A frequência mínima estipulada para sua realização pode depender do órgão regulador do país e/ou da empresa responsável, mas de forma geral, os testes são feitos mensalmente.

O teste de produção de determinado poço é conduzido desviando a sua produção para um separador de teste ou um medidor multifásico, mantendo-se as condições normais de produção. No caso desse trabalho, o teste é realizado em um separador de teste trifásico localizado na plataforma. São então medidas, a partir das correntes separadas, as vazões de água, óleo e gás do poço em análise, como é mostrado na Figura

1. Além disso, outras importantes medidas características da produção são registradas. Ao final do teste, as informações obtidas são registradas em boletins de testes. Nestes, as variáveis do processo são anotadas por hora. O valor final de cada variável é obtido pela média desses registros horários. A Tabela 1 mostra um exemplo simplificado de algumas variáveis registradas em um boletim.

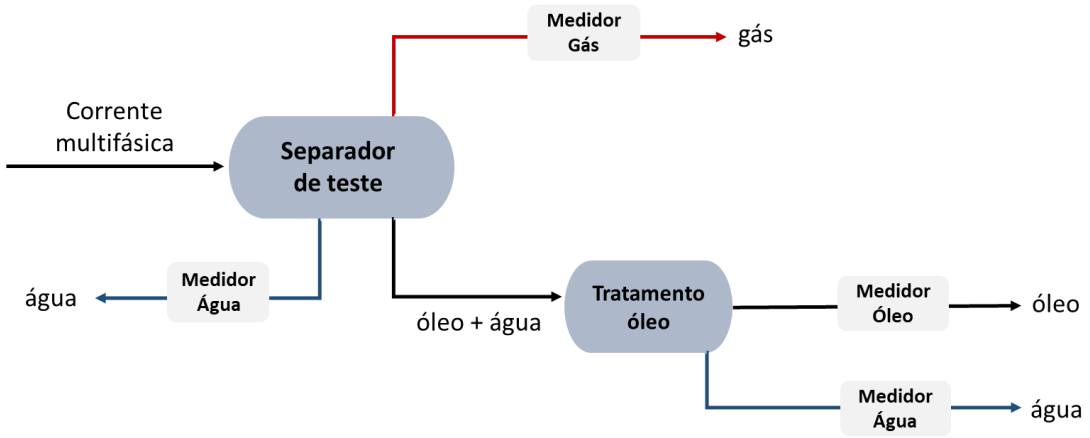


Figura 1: Esquema simplificado de um separador de teste.

Tabela 1: Exemplo das medidas obtidas por um teste de produção.

Data/Hora	***Fundo do Poço***		***Cabeça do Poço***		***Separador de Teste***					
					Separação		***Saída de Óleo***		***Saída de Água***	***Saída de Gás***
	Pressão	Temp.	Pressão	Temp.	Pressão	Temp.	Vazão de Óleo	Vazão de Água	Vazão de Água	Vazão de Gás
	kgf/cm²	*°C*	*kgf/cm²*	*°C*	*kgf/cm²*	*°C*	*m³/h*	*m³/h*	*m³/h*	*m³/h*
01/08 19:00	170	55	109	47	9	69	32	20	20	11213
01/08 20:00	170	55	110	47	9	73	32	20	21	11406
01/08 21:00	170	55	109	47	9	74	32	20	21	11155
01/08 22:00	170	55	110	47	9	74	31	20	21	11296
Média	170	55	110	47	9	72	32	20	21	11268

Após o término do teste, todas essas informações coletadas pelo boletim serão examinadas por uma equipe de especialistas para verificar a validade do teste. Essa validação é um processo complexo e subjetivo. De forma geral, testes são invalidados em caso de alterações muito grandes em relação a boletins anteriores, caso seja verificado que o teste não foi conduzido de forma correta por problemas operacionais ou ainda caso o poço não esteja estabilizado durante a realização do teste.

Um fator que aumenta a complexidade de validação de testes é a existência de erros e incertezas oriundas do processo de medição e dos fenômenos estocásticos existentes na produção de petróleo. É difícil definir exatamente quando vão ocorrer mudanças nas condições do poço. Isto gera uma complicação na validação do novo teste. Por exemplo, um valor destoante de fração de água produzida, em relação aos testes anteriores, pode significar que ocorreu algum erro durante a realização do teste, e este realmente deveria ser invalidado, ou foi simplesmente uma mudança de comportamento no poço que ocasionou variação do valor de fração de água.

Além disso, as incertezas e erros podem também estar associados a diversos problemas que ocorrem durante a realização dos testes, como aparelhagem fora do intervalo de operação adequado, medições incorretas de volumes e outros parâmetros, emulsões durante a produção e mau funcionamento de *softwares*. Estes problemas podem acarretar em testes de produção não representativos. Entretanto, de forma geral, esses problemas só são identificados após a finalização do teste, quando engenheiros encarregados irão analisar os dados obtidos e validar ou não o teste de acordo com as informações contidas.

A atividade de testes de produção em um campo ainda enfrenta problemas na quantidade limitada de testes realizados. Em um cenário ideal, um separador de teste ou um medidor multifásico seria instalado para cada poço, de forma que testes fossem frequentemente realizados. Entretanto, isto não é viável em vista dos altos custos associados e por questões de limitação de espaço. Assim, geralmente, o separador de teste é compartilhado por um conjunto de poços e os poços são testados, de forma individual, periodicamente. Como consequência, em campos com muitos poços, pode-se passar um longo período entre a realização de testes em cada poço.

Alguns trabalhos na literatura (BRUNI et al., 2003; OLSEN; NORDTVEDT, 2006) propõem o uso de simuladores de escoamento para estimar valores esperados para melhorar o processo de validação de testes de produção. Testes de produção são também responsáveis por adquirir parâmetros que serão utilizados para calibrar modelos de produção e reservatórios. Mas, esses próprios modelos computacionais podem ser utilizados na validação de testes de produção. No entanto, muitas vezes, as incertezas e erros associados à atividade de produção não conseguem ser captadas pelos modelos. Além disso, geralmente esses modelos demandam muito tempo computacional em vista do grande número de variáveis e da complexidade do problema. Dessa forma, a

utilização de técnicas de mineração de dados, cujos modelos são baseados no conjunto de dados disponíveis, se torna um caminho em potencial para ser explorado, inclusive em validação de testes de produção.

A indústria de petróleo nos seus diferentes segmentos, cada vez mais, está buscando a utilização de métodos de mineração de dados e técnicas inteligentes para resolução dos seus problemas de alta complexidade. Projetos pilotos em alguns campos de produção, como o campo Sabriya, no Oriente Médio, estão sendo conduzidos na busca de operações mais inteligentes, integradas e de alto desempenho. Dessa forma, se considera tais práticas como um caminho a ser seguido pelos campos brasileiros, na busca de operações de produção mais eficientes e de alta performance.

Em vista do que foi apresentado, uma ferramenta de análise de variáveis e validação de teste, a partir das medidas obtidas durante a realização do mesmo, colaboraria com o processo de análise e de tomada de decisão. Além disso, permitiria identificar possíveis problemas enfrentados, e se fosse possível, corrigi-los em tempo hábil. Em caso de um fator irreparável, o teste poderia ser finalizado e invalidado antes de cumprir certo número de horas esperadas. Assim como, se as variáveis do teste estivessem dentro de um intervalo esperado, se teria mais confiabilidade dos resultados obtidos durante os testes de produção.

1.2 Objetivos

O objetivo geral desse trabalho é desenvolver uma ferramenta baseada em mineração de dados, para validação de testes de produção, que possa avaliar as variáveis obtidas durante a realização do teste e identificar se estas estão de acordo com o comportamento esperado.

Como objetivos específicos pretende-se definir um modelo de classificação que consiga rotular os dados como válidos ou não válidos; identificar possíveis problemas inerentes ao processo de testes de produção, auxiliando o operador responsável na realização do teste e colaborando na obtenção de testes com melhor qualidade. Além disso, espera-se desenvolver modelos de regressão para analisar as vazões de óleo, água e gás, que são as variáveis-alvos dos testes de produção, assim como intervalos de predição robustos, de forma a se considerar as incertezas presentes no processo de produção.

1.3 Estrutura do Trabalho

Esta dissertação está dividida em oito capítulos. Neste capítulo foi contextualizado o tema a ser estudado e apresentados os objetivos do trabalho.

No segundo capítulo é definido o problema a ser estudado.

No terceiro capítulo é mostrada a parte teórica da atividade dos testes de produção, assim como trabalhos na literatura que relatam sobre a validação de testes de produção.

No quarto capítulo é feita uma contextualização sobre mineração de dados. Além disso, são explicados os modelos de regressão e classificação utilizados neste trabalho. Ainda neste capítulo é feita uma revisão bibliográfica sobre trabalhos da literatura que relatam o uso de métodos de mineração de dados e de novas tecnologias inteligentes aplicadas ao processo de produção de petróleo.

O procedimento metodológico proposto para resolver o problema é mostrado no capítulo 5.

No capítulo 6 é feita a experimentação do método desenvolvido através de um estudo de caso.

No capítulo 7 são mostradas as conclusões obtidas e sugestões para trabalhos futuros.

Finalmente no capítulo 8 são mostradas as referências bibliográficas utilizadas como suporte para este trabalho.

2 DEFINIÇÃO DO PROBLEMA

De forma geral, somente após a finalização do teste de produção que os especialistas irão avaliar as informações obtidas e decidir sobre sua validação. Entretanto, realizar um teste de produção demanda muito tempo e esforço, além de que as unidades de produção estão normalmente ligadas a muitos poços e um único separador de teste de produção é utilizado para atender essa demanda. Isso gera uma limitação na realização contínua de testes em cada poço. Assim, é muito prejudicial para o bom desenvolvimento da produção, um determinado poço passar por todo o processo de teste produção para no final obter a informação de que este não teve bom desempenho. Testes de produção inválidos devem sempre ser evitados com uma preparação correta do teste. Mas, mesmo com boas práticas de realização, estes ainda podem ser invalidados, em função de problemas operacionais e incertezas da produção.

Dessa forma, o que esse trabalho pretende responder é como desenvolver, a partir do conjunto de variáveis de operação disponíveis, uma ferramenta eficiente que seja capaz de validar, em tempo real, testes de produção periódicos. Trata-se de dois problemas acoplados. Um mais geral, que procura criar modelos preditivos para classificar um teste e analisar o comportamento das variáveis de forma automática, a partir de um histórico de produção das variáveis procedentes de testes anteriores. E um outro problema que é mais pontual, que busca, durante a realização do teste, determinar se existe um ponto ótimo que aquele teste é considerado válido ou não. Ou seja, a partir dos registros de variáveis de produção que são obtidas por hora, identificar se é possível determinar se um teste que será invalidado, já começou com um comportamento inválido, ou se ele modifica seu comportamento ao longo do tempo, começando na categoria inválido e passando durante o teste para a categoria válido. Além disso, determinar a partir de qual horário, desde o início do teste, que este pode ser definido, com alto nível de confiabilidade, como válido ou inválido.

Este trabalho tem como foco o primeiro nível do problema. Ou seja, em como criar modelos preditivos eficientes, a partir do histórico de dados de teste de produção obtidos ao longo da vida produtiva do poço. A partir da resposta obtida por esta parte, uma análise apurada no segundo problema poderá ser realizada. Uma primeira dificuldade que surge nessa temática é em como lidar com as incertezas existentes no processo de produção de petróleo. É necessário determinar os diferentes resultados

possíveis que podem ser encontrados para a variável analisada, de forma que o problema precisa ser tratado com uma abordagem estocástica.

Outro aspecto importante para este problema é determinar quais as variáveis e parâmetros que serão entradas dos modelos e quais as saídas desejadas. Dentre os dados disponíveis para realização do trabalho, se tem as medidas diretas registradas de hora em hora, em que o resultado final da variável no teste é dado pela média desses registros horários. Desse grupo, fazem parte as medidas de pressões, temperaturas e vazões.

Logo após a saída do reservatório, no fundo poço, encontram-se medidores de pressão e temperatura do fluido produzido. O medidor de pressão e temperatura deste ponto é chamado PDG (*Pressure Downhole Gauge*). O segundo ponto de pressão e temperatura disponível nesse trabalho vem do medidor TPT (do inglês, *Tree Pressure and Temperature*). O Medidor TPT está localizado na árvore de natal, na cabeça do poço. Na árvore está um conjunto de válvulas e equipamentos que controlam o fluxo de petróleo e gás. Ainda no poço, existem as válvulas de injeção de *gas lift*, que é o método de elevação artificial utilizado nos poços desse trabalho. Para este método, existem as medidas de pressão de injeção do *gas lift* no poço e a vazão de injeção, que são duas variáveis com registros horários também disponíveis para serem utilizadas neste trabalho. Ao sair do poço, o fluido é encaminhado por linhas de produção até a unidade de produção. Nesta unidade também existem medidores de pressão e temperatura, que também estão nos registros e são considerados. Na plataforma ainda existem os medidores de vazão bruta, que determinam a produção de líquido do fluido produzido. A esquematização do que foi exposto neste parágrafo pode ser visto na Figura 2.

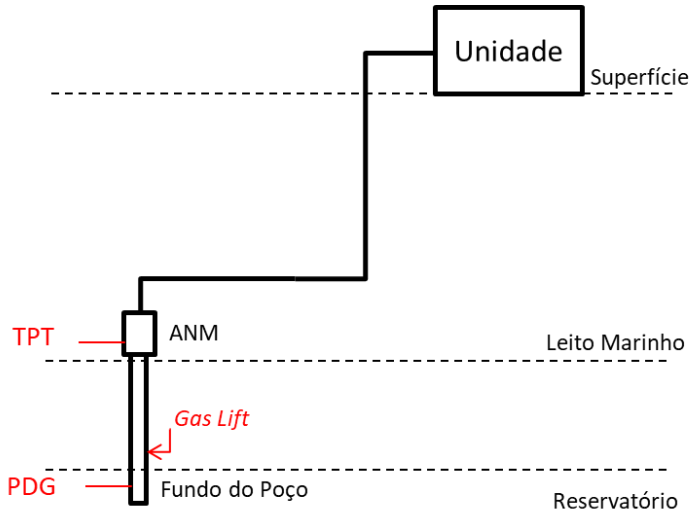


Figura 2: Esquematização do sistema de produção. Fonte: Elaboração Própria.

A maior parte das medidas listadas podem ser obtidas diariamente, e não somente quando teste de produção são conduzidos. Entretanto, medidas de vazão de óleo, água e gás, assim como a pressão e temperatura de separação do fluido no separador de teste são somente coletadas na realização do teste de produção.

Além das medições horárias listadas, existem outras variáveis que são muito utilizadas em análises de campos e poços de petróleo e que também se tem dados disponíveis. Nesta categoria entram as informações de fração de água produzida (BSW), razão gás-óleo (RGO) e razão gás-líquido (RGLI). Estas variáveis são obtidas por meio de cálculos usando as médias das variáveis de análise horária.

BSW é a fração de água produzida, e corresponde a porcentagem de volume de água em relação ao volume de líquido produzido. Nos testes de produção, esta medida é calculada através da razão entre a média obtida pela vazão de água e a média da vazão de óleo. O valor de BSW pode ser obtido também através de amostras coletadas na plataforma. Assim, é possível se ter as medidas de BSW diariamente, e não somente quando testes de produção ocorrem. A razão gás-óleo (RGO) do poço é a fração de gás pela fração de óleo, e depende das vazões de gás e óleo e razão gás-líquido (RGLI) é a medida da razão de gás e líquido.

Por fim, existem medidas que não são obtidas pelos testes de produção, mas que são muito importantes para a atividade da produção e que poderiam ser adicionadas a análise. Neste caso, compõem esse grupo o índice de produtividade do poço (IP), pressão estática do reservatório, razão de solubilidade óleo-gás (RS) e fator de

encolhimento do óleo (FE). Essas são medidas do reservatório obtidas por outros tipos de testes.

As variáveis disponíveis são uma parte de todas variáveis que participam do processo de produção de petróleo. Mas ainda assim, são muitas variáveis e parâmetros a serem analisados. Como o objetivo principal dos testes de produção periódicos é a obtenção das vazões de óleo, água e gás do poço, estas serão as variáveis dependentes a serem analisadas. Para este caso de previsão das vazões, os dados das outras variáveis serão considerados as entradas do modelo de previsão.

Outro ponto importante do trabalho é determinar se um teste é válido ou não válido, a partir do conjunto de registros disponíveis. Neste caso, todas as variáveis podem fazer parte da entrada do problema, inclusive as vazões de óleo, água e gás.

Dessa forma, para os dois problemas descritos, classificação do teste e previsão das variáveis, um primeiro problema é a seleção das variáveis a serem consideradas. Neste ponto, é necessário identificar se todas essas variáveis influenciam nos testes de produção, ou quais delas tem um impacto maior. Além disso, é ainda necessário avaliar quais variáveis e parâmetros que serão entradas dos modelos para as saídas desejadas.

Finalmente, uma última questão levantada necessária nesse trabalho é como desenvolver uma abordagem estocástica para lidar com as incertezas inerentes ao processo de produção de petróleo, e obter intervalos de previsão robustos de valores esperados das variáveis de saída analisadas.

3 REVISÃO BIBLIOGRÁFICA DE TESTES DE PRODUÇÃO DE PETRÓLEO

O trabalho proposto se funde na utilização de ferramentas de mineração de dados e na automatização do processo operacional de validação de testes de produção de petróleo. Assim, o objetivo deste capítulo é abordar como os testes de produção são feitos e mostrar os trabalhos na literatura que tratam sobre validação de testes de produção de petróleo, assim como sua automatização. Como alguns trabalhos a serem mostrados relatam sobre modelos físicos do processo de produção de petróleo que requerem um conhecimento prévio sobre atividades de elevação e escoamento de petróleo, neste capítulo é ainda descrito resumidamente os principais conceitos de produção de petróleo.

Nos itens que seguem, primeiramente é dada uma visão geral sobre testes de produção de petróleo, após isso, conceitos básicos de produção de petróleo são expostos e finalmente, são mostrados trabalhos na literatura sobre validação dos testes de produção de petróleo. Antes, é importante mencionar que ainda é incipiente os estudos sobre validação de testes de produção, especialmente os que se referem a sua automatização.

3.1 Testes de Produção de Petróleo

Segundo ALLEN e ROBERTS (1982), os objetivos de testes de produção podem variar desde a determinação da quantidade de fluidos produzidos até a determinação dos parâmetros do reservatório e suas heterogeneidades. De forma geral, os testes de produção de poços de óleo ou gás são classificados, de acordo com suas funções, como testes de produção periódicos, testes de produtividade ou testes de pressão transiente. No caso desta dissertação, os testes de produção analisados são do tipo periódicos.

Os testes de produção periódicos têm a função da determinação das quantidades relativas de vazões de óleo, água e gás produzidos pelo poço em condições normais. Normalidade nesse contexto significa que o poço deve estar produzindo sua quantidade média normal de óleo, água e gás, incluindo as vazões otimizadas de injeção do método de elevação artificial de *gas lift*. Ainda neste aspecto é importante que o poço esteja

estabilizado antes da realização do teste de produção, de forma que problemas, como emulsões, devem estar controlados antes do início do teste.

Com as informações obtidas pelos testes de produção, é possível alocar a produção total dos poços em produção no campo analisado, e cumprir a regulação que exige relatórios periódicos dos poços produzidos. No caso brasileiro, por exemplo, a Agência Nacional de Petróleo, Gás Natural e Biocombustíveis (ANP) exige a elaboração de boletins de produção.

A realização dos testes de produção é também fundamental para as atividades de operação, pois permite identificar como estão as condições correntes do poço. Mudanças inesperadas como produção exagerada de água e gás podem indicar problemas no reservatório ou no poço. Uma queda abrupta da produção pode estar relacionada, por exemplo, a problemas na elevação artificial, ou produção de areia (ALLEN; ROBERTS, 1982).

O teste de produção periódico pode ser feito através de um medidor multifásico acoplado ao poço ou por separador de teste. O último é mais usual, especialmente para campos *offshore*, por questões econômicas. Geralmente, na unidade de produção existe apenas uma planta de teste a ser compartilhada pelos poços produzidos pela unidade. Para realizar o teste de produção em determinado poço, este é alinhado à planta de teste, e após sua produção for estabilizada, as medições começam ser armazenadas. O número de testes e sua frequência será dependente do número de poços que estão associados a mesma planta de separação (SÆTEN, 2015). A Figura 3 mostra um esquema de um sistema de produção com a planta de teste. Os quatro poços da figura estão associados a planta de produção e a de teste, por meio de válvulas. Quando determinado poço entrar em teste, este é alinhado ao separador, onde as medidas serão aferidas.

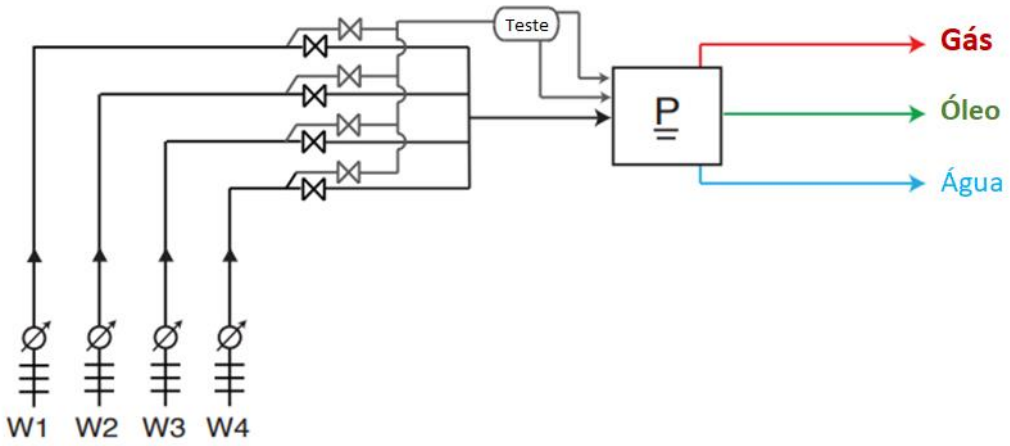


Figura 3: Esquema do sistema de produção com a planta de teste associada. Fonte: Adaptado de SÆTEN (2015).

3.2 Conceitos Básicos de Escoamento de Petróleo

A produção de um campo de petróleo depende da elevação e do escoamento dos fluidos desde o reservatório até a unidade de produção. No começo da produção, geralmente o poço produz com a própria pressão do reservatório, ou elevação natural. Os mecanismos básicos de produção, os quais determinam o comportamento dos poços durante a vida produtiva do poço são: gás em solução, influxo de água e capa de gás (ROSA; XAVIER; CARVALHO, 2006). À medida que a produção do reservatório acontece, sua energia diminui, e a pressão no fundo do poço passa a ser insuficiente para elevar os fluidos até a unidade de produção, tornando-se necessária a utilização dos métodos artificiais de elevação para a produção do campo. Além disso, alguns reservatórios, mesmo no início da vida produtiva, possuem pressão relativamente baixa, nestes casos também são aplicados os métodos artificiais. Os métodos artificiais fornecem trabalho ao sistema de produção, e atuam, por exemplo, reduzindo a densidade do fluido para que este possa ser produzido. Os métodos mais comuns são: *Gas lift*, contínuo e intermitente; Bombeio centrífugo submerso, BCS; Bombeio mecânico com hastes, BM; Bombeio de cavidades progressivas, BCP (THOMAS, 2001).

Para garantir a elevação e o escoamento de petróleo, os fluidos precisam passar pelas etapas de fluxo no meio poroso, na coluna de produção e através da linha de produção ou restrições. Além disso, é necessário que tenham energia suficiente para

superar as perdas de carga. O perfil da pressão de ao longo dessas etapas é mostrado na Figura 4.

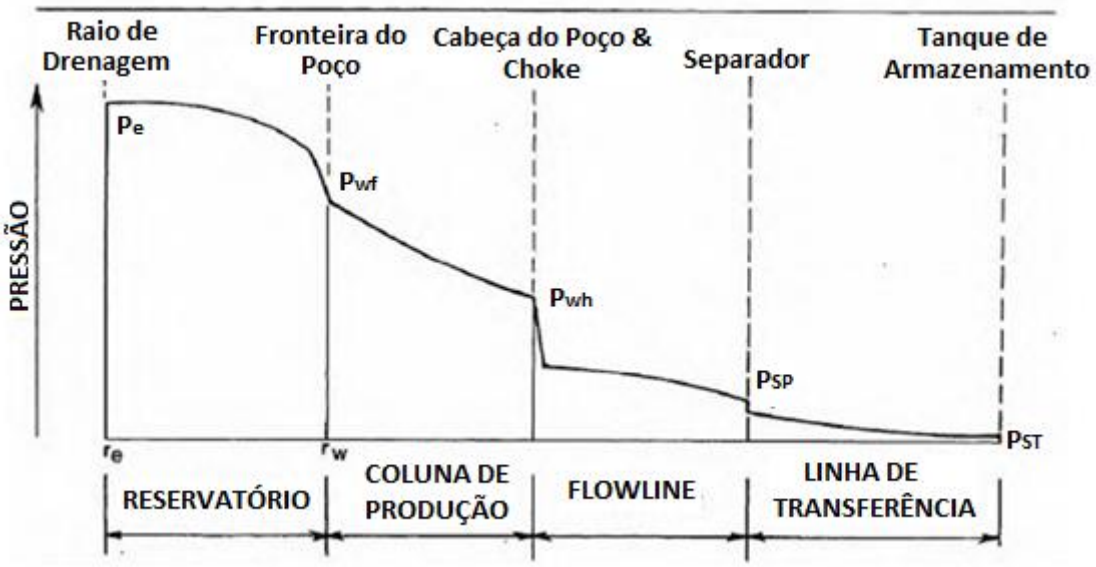


Figura 4: Perfil da Pressão na Elevação e Escoamento de Petróleo (Fonte: Adaptado de LYONS, 1996).

3.2.1 Fluxo no meio poroso

O fluxo no meio poroso corresponde ao fluxo de fluidos dentro do reservatório e é representado pela curva de pressão disponível no fundo do poço, ou *Inflow Performance Relationship* (IPR) (GILBERT, 1954). Esta curva indica a relação entre a pressão de fluxo no fundo do poço e a vazão de líquido no meio poroso. Quando a pressão do reservatório é maior que a de saturação, todo gás está dissolvido no líquido e a IPR é linear, conforme mostrado na Figura 5 (THOMAS, 2001). A equação característica da IPR linear é dada pela Equação 1:

$$\mathrm{P_{wf}} = \mathrm{P_e} - \frac{\mathrm{q}}{\mathrm{IP}} \qquad \text{Equação 1}$$

Em que P_e é a pressão estática do reservatório, P_{wf} é a pressão de fluxo no fundo do poço e IP é o índice de produtividade do poço, e q é a vazão de produção do líquido.

Entretanto, para os casos em que a pressão do reservatório é menor que a pressão de saturação, o índice de produtividade do poço varia com a pressão do fluxo no fundo do poço, não sendo mais aceita a representação linear da IPR. Para esses casos, existem diferentes formas de representar as curvas de performance do poço, entre elas destaca-se

o modelo de Vogel (1968), Figura 6, adequado para vários tipos poços (THOMAS, 2001). A equação de Vogel é dada pela Equação 2.

$$\frac{q}{q_{máx}} = 1 - 0{,}2\frac{P_{wf}}{P_e} - 0{,}8\left(\frac{P_{wf}}{P_e}\right)^2 \qquad \text{Equação 2}$$

Em que q é a vazão de produção do poço prevista pelo modelo e $q_{máx}$ é a vazão máxima teórica de produção do poço.

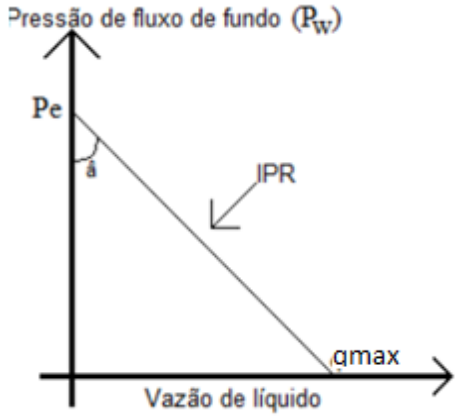


Figura 5: Modelo de IPR Linear (Fonte: Adaptado de THOMAS, 2001).

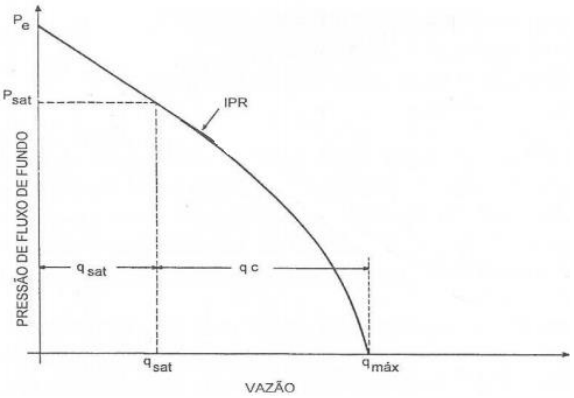


Figura 6: Curva de IPR - Modelo de Vogel. (Fonte: Adaptado de ROSSI, 2004).

3.2.2 Fluxo na coluna de produção e nas linhas de produção

Os fluidos devem ter pressão suficiente para percorrer a coluna de produção, as linhas e *risers*. Para ocorrer produção, a pressão disponível no fundo do poço deve ser capaz de vencer as perdas de carga da coluna hidrostática de fluido, perdas por fricção na coluna de produção, linhas e *risers*, perdas nas restrições de superfície e sub-superfície e chegar no separador com a pressão de separação requerida. Nessas etapas, o comportamento do escoamento dos fluidos nas tubulações é determinado pela equação do gradiente de pressão ou equação do balanço de energia mecânica, que é resultado das leis de Conservação de Massa, Momento e Energia. Assim, a equação do gradiente de pressão é formada por três componentes: elevação, fricção e aceleração, como mostrado na Equação 3 (BRILL; MUKHERJEE, 1999).

$$\frac{dp}{dL} = \left(\frac{dp}{dL}\right)_{el} + \left(\frac{dp}{dL}\right)_{f} + \left(\frac{dp}{dL}\right)_{ac} \qquad \text{Equação 3}$$

A primeira componente da Equação 3 refere-se à perda de carga por elevação ou coluna hidrostática. Normalmente é predominante na coluna de produção e representa entre 80 a 95% da perda de carga total. A segunda corresponde à parcela de perda por fricção na parede do duto e contribui de 5 a 20% do gradiente de pressão na coluna. A terceira parcela é a aceleração, relacionada à variação da velocidade. Normalmente é negligenciada nos cálculos, sendo somente considerada nos casos com alta velocidade de escoamento, como o caso de poços de *gas lift* perto da superfície (BRILL; MUKHERJEE, 1999).

A maior parte da produção de petróleo estão em condições de fluxo multifásico, em que geralmente tem-se óleo com gás em solução, gás livre e água. São fluidos com propriedades físicas diferentes escoando simultaneamente em um mesmo duto, o que torna a análise muito mais complexa, pois existe uma ampla gama de padrões de fluxo possíveis, dificultando assim os cálculos de gradiente de pressão. Padrão de fluxo refere-se à distribuição de cada fase em um duto relativa à outra fase e depende principalmente da velocidade do gás e do líquido e da relação entre gás e líquido. A predição de padrões de fluxo em dutos horizontais, recorrente nas linhas de produção, é ainda mais complexa que nos dutos verticais, pois as fases tendem a se separar em função da diferença de densidade (BEGGS, 2003).

Em função da dificuldade de se calcular o gradiente de pressão no escoamento multifásico, os simuladores de fluxo multifásico em tubulações são muito utilizados. Estes simuladores utilizam correlações empíricas. Estas podem ser divididas em três categorias, A, B e C, de acordo com a consideração de padrão de fluxo e escorregamento entre as fases (THOMAS, 2001). Dentre as correlações mais usadas, destaca-se o método de BEGGS & BRILL (1973), da categoria C, que pode ser utilizado para dutos com qualquer ângulo de inclinação.

Como o fluido segue desde o reservatório até a chegada na unidade de processamento, e ao longo desse percurso ocorre perda de carga, é importante analisar a pressão em um determinado ponto do escoamento. Para isso, normalmente é realizada a análise nodal, em que se divide um sistema de produção em partes, e equações matemáticas são aplicadas em cada uma das partes de forma individual. Assim, escolhe-se um nó de análise, e para cada nó do sistema, duas curvas de pressão em função da vazão são traçadas, uma de pressão disponível e outra de pressão requerida (BEGGS, 2003).

3.3 Trabalhos Relacionados a Validação de Testes de Produção de Petróleo

BRUNI et al. (2003) descrevem um sistema construído para monitorar e garantir a validação de testes de produção, através da ferramenta PROMO (PROduction Management and Optimization), ferramenta que integra *softwares* e fluxos de trabalho. No sistema criado, uma interface faz a comunicação entre dois *softwares*, um responsável pela análise, que faz os cálculos de desempenho e a modelagem da produção utilizando o último teste válido, e outro responsável pela validação, que recebe os resultados do primeiro *software* e verifica a diferença entre o resultado modelado e o novo teste sob análise, validando ou não o teste. Quando um novo teste é aprovado e validado, a curva IPR (do inglês, *Inflow Performance Relationship*) é automaticamente calibrada.

Uma das principais vantagens apresentada pelo trabalho de BRUNI et al. (2003) é a integração entre as ferramentas que fornecem testes de produção mais confiáveis e com menor tempo, pois o tempo consumido na manipulação de dados foi eliminado. Além disso, a possibilidade de visualizar o novo dado testado com o histórico do sistema permite que o engenheiro possa identificar com maior facilidade quais poços não estão se comportando conforme o esperado.

O artigo de OLSEN e NORDTVEDT (2006) mostra uma ferramenta automatizada de análise de testes de poços de reservatórios. Dos testes, são obtidas séries temporais de parâmetros característicos de reservatório, como efeito de película (*skin factor*), permeabilidade, pressão do reservatório e área drenada. Na ferramenta desenvolvida, primeiramente os dados são filtrados e compactado em um módulo de filtragem automatizada, utilizando o filtro de *Wavelet*. O objetivo deste módulo é melhorar o desempenho computacional e detectar eventos automaticamente.

Após a preparação, os dados são encaminhados para o segundo módulo de análise de testes de poços. Neste, utiliza-se regressão não linear com intervalos de confiança. A otimização é feita utilizando o algoritmo *Levenberg-Marquardt* com derivadas analíticas. Para detectar variações nas séries temporais dos parâmetros estimados, são analisadas as mudanças no valor médio e na variância do parâmetro, através de um teste de hipótese com um intervalo de confiança. A detecção de mudança na média é baseada no teste t de *Student* que fornece a probabilidade de duas amostras

terem medidas diferentes e ainda apresentarem a mesma média ou variância. Segundo os autores, o teste de hipótese pode ser utilizado para detectar mudanças em parâmetros e distinguir entre testes de poços errados e testes de poços com mudanças de parâmetros. Os resultados obtidos em análises feitas utilizando dados sintéticos foram satisfatórios.

CRAMER et al. (2006) mostram uma ferramenta de instrumentação automatizada desenvolvida pela Shell, para a otimização e automatização dos testes de produção de poços. A validação dos testes está entre os módulos disponíveis pela ferramenta. O módulo criado utiliza uma combinação de modelos baseados em dados e reconhecimento de padrões que permite identificar o tempo ótimo de estabilização do poço, duração do teste e qualidade dos resultados. O modelo construído utiliza dados de vazões do separador e de pressões obtidos em tempo real. O método desenvolvido realiza predições dos valores esperados das medidas de vazões ao longo do teste, levando em consideração o histórico de produção. Se as medidas obtidas estão de acordo com as do modelo, o tempo de parada do teste é atingido e este é automaticamente validado.

RODRIGUEZ et al. (2013) relatam sobre um projeto piloto de operações de campos de óleo digitais inteligentes (iDOF) que está sendo conduzido no campo Sabriya, no Oriente Médio. Este programa propõe uma nova abordagem para a indústria de petróleo que integra ferramentas tradicionais de reservatório e produção, ferramentas estatísticas, agentes inteligentes e simulação numérica, com o objetivo final de uma análise rápida de dados e compartilhamento do conhecimento de diferentes áreas para uma melhor tomada de decisão.

O artigo ainda mostra como a arquitetura do fluxo de trabalho automatizado é dividida. Neste, são utilizadas ferramentas analíticas para entender o histórico de produção, ferramentas estatísticas para monitorar a produção atual em tempo real e são usados agentes de inteligência artificial para previsão da produção em curto tempo. Em ferramentas analíticas, são usadas análise nodal, curva de declínio, medidor virtual e simulação numérica. Em ferramentas estatísticas, é utilizada correlação linear, sinalização de alarme e alerta, filtragem e condicionamento de dados, análises de Monte-Carlo, histogramas e gráficos de Pareto. Finalmente, em agentes inteligentes, utilizam-se ferramentas de reconhecimento de padrões, redes neurais, lógica Fuzzy e análise da causa raiz (RCA).

Do projeto piloto de operações de campos de óleo digitais inteligentes (iDOF) relatado, algumas linhas de pesquisa foram criadas. Destas pesquisas, um dos módulos desenvolvidos se refere diretamente a validação de testes. O trabalho de CULLICK et al. (2013) relata sobre este módulo, denominado Avaliação de Desempenho de Poço (WPE). O objetivo principal do módulo é acelerar, priorizar e simplificar o processo de testes de poços e suas análises relativas. O artigo menciona práticas para supervisionar testes de poços em tempo real e obter respostas rápidas a comportamento anômalos identificados no poço. Contém módulos de filtragem e limpeza dos dados de testes e medidores multifásicos que captam informações instantâneas dos testes.

Ainda no módulo WPE, CULLICK et al. (2013) mostram uma forma eficiente de otimizar o processo de testes de produção, que consiste em ajustar a válvula *choke*, válvula responsável pelo controle de fluxo, em três posições diferentes ao longo do tempo do teste. Assim, medidas de vazão de líquido e pressão na cabeça do poço (THP) são obtidas em três períodos distintos do teste. Para cada período, é calculado o valor médio de THP e vazão de líquido. Com esses três pontos, é viável estimar a pressão de reservatório e vazão máxima de líquido, através de extrapolação linear. Utilizando esses valores como estimadores iniciais na equação de Vogel, várias iterações são conduzidas até se obter o menor coeficiente de correlação. Se este coeficiente ainda for maior que 0,9, o teste é aceito e os valores médios dos dados obtidos são transferidos para seção de modelagem. Apesar de não se ter condições de aplicar esse método neste trabalho, acredita-se que é um processo operacional útil para companhias de petróleo e que poderia ser estudado sua aplicabilidade durante a realização de testes de produção de poços para obtenção de melhores resultados.

Técnicas analíticas para obtenção de operações mais inteligentes estão sendo cada vez mais utilizadas na indústria de petróleo. BRAVO et al. (2014) apresentam uma análise das principais técnicas analíticas aplicadas em atividades de produção de óleo e gás, e como essas ferramentas podem ser utilizadas para apoiar operações inteligentes. Uma das aplicações mencionadas no artigo é a validação de testes e atualização do modelo do poço. Nestas atividades, as ferramentas analíticas são utilizadas para melhorar a qualidade dos resultados obtidos pelos testes e minimizar sua duração, maximizando assim o uso do separador de teste. O artigo ainda apresenta um fluxo de trabalho para realização dos testes que se divide nas operações de tratamento dos dados, validação de testes, comparação do desempenho do teste com um modelo esperado,

análise de incompatibilidade dos resultados e ajuste de parâmetros do modelo de produção do poço. Técnicas Analíticas podem ser integradas a essas etapas para obtenção de melhores resultados.

Ainda no artigo de BRAVO et al. (2014), os autores mencionam que ferramentas analíticas podem ser aplicadas no monitoramento da produção para identificar eventos que ocorrem no processo. Podem ser utilizadas, por exemplo, redes neurais, lógica *Fuzzy* e redes bayesianas. Outro emprego de técnicas analíticas está na medição virtual de variáveis, em que ferramentas baseadas em redes neurais são utilizadas para modelar o comportamento de poços de óleo e prever vazões instantâneas e futuras de óleo. Apesar de citar diferentes atividades que as ferramentas analíticas poderiam ser aplicadas na área de produção de petróleo, o trabalho não detalha o procedimento para essas aplicações.

NASER E ZAINAL (2015) descrevem o processo de melhoria dos testes de produção no campo de petróleo Bahrain ao se adotar novas práticas. Quando um separador de teste de duas fases foi substituído por um de três fases, inicialmente, poucos testes eram validados pelos engenheiros de produção. Foi então desenvolvido um plano de trabalho para melhorar a validação dos testes dos poços. Dentre as medidas tomadas, destacam-se a instrumentação e automatização do processo de testes de poços e o desenvolvimento de indicadores-chave de desempenho (do inglês, *Key Performance Indicator*, KPI).

Dentre os indicadores analisados em NASER E ZAINAL (2015), a categoria que tem maior importância para este trabalho são as razões gerais para qualidade insatisfatória dos testes. As causas incluem medições incorretas de volume de líquido, da fração de água produzida (*water cut*), volume de gás e questões operacionais. Esses problemas eram decorrentes de um intervalo inadequado dos níveis do separador, emulsões durante a realização dos testes e erros nos *softwares*. As informações apresentadas no artigo colaboram no processo de mapear os possíveis problemas que podem ocorrer durante a realização dos testes. Vale ressaltar que após as medidas de melhoria tomadas, a validação dos testes de produção aumentou de 30% para 90% no campo de Bahrain.

O trabalho de RAO; DAVID (2015) é uma referência pertinente para identificar como são as práticas atuais de testes de produção de poços e seus desafios. Dentre o que foi exposto, o artigo menciona o aspecto crítico que é a validação de testes no processo.

Como práticas correntes, a validação é feita comparando os valores obtidos com as tendências do histórico de dados. Além disso, para lidar com as mudanças de vazões, de propriedades dos fluidos e do reservatório, pode-se também comparar as vazões obtidas durante o teste com vazões estimadas por modelos de poços. No entanto, esses modelos devem estar calibrados e atualizados com parâmetros recentes. É então que o artigo propõe a criação de uma ferramenta automatizada de validação de dados de testes que integre o processo dos testes com o uso contínuo do modelo de poços.

Além disso, a ferramenta também pretende lidar com o longo período entre dois testes consecutivos, que é um segundo grande desafio da atividade de testes. Para fornecer vazões contínuas entre o intervalo de dois testes consecutivos, são usadas medições de vazões virtuais conjuntamente com resultados de modelos de poços. Em medições de vazões virtuais, é usado a técnica de modelos de redes neurais que utiliza histórico de testes de poços para gerar estimativas de vazões dos poços. A estrutura digital de testes de produção integra a área de reservatórios e de produção de campos de petróleo, permitindo uma produção otimizada e melhor gerenciamento do campo.

A seguir são mostrados um resumo com os trabalhos expostos anteriormente neste item.

Tabela 2: Resumo dos trabalhos de validação de testes de produção de petróleo.

Trabalhos	**Validação de Testes**	**Produção**	**Método**
BRUNI et al. (2003)	✓		Ferramenta que integra softwares e fluxos de trabalho.
OLSEN e NORDTVEDT (2006)	✓		Filtro de Wavelt; Levenberg-Marquard; teste t-Student.
CRAMER et al. (2006)	✓		Modelos baseados em dados e reconhecimento de padrões.
CULLICK et al. (2013)	✓		Filtragem, limpeza dos dados de testes; melhoria do processo operacional.
BRAVO et al. (2014)	✓	✓	Fluxo de trabalho, redes neurais, lógica *Fuzzy* e redes bayesianas.
NASER e ZAINAL (2015)	✓		Automatização de testes de poços e o desenvolvimento de KPI.
RAO e DAVID (2015)	✓	✓	Integração de testes com modelos; Redes Neurais.

4 REVISÃO BIBLIOGRÁFICA DE MINERAÇÃO DE DADOS

Segundo HAN et al. (2012), o conceito de mineração de dados é tratado ou como um sinônimo para o processo de Descoberta de Conhecimento em Base de Dados (do inglês, *Knowledge Discovery from Data* – KDD), ou apenas como uma etapa essencial desse processo. A metodologia KDD é mostrada de forma simplificada na Figura 7.

Figura 7: Processo KDD. Fonte: Figura adaptada de FAYYAD; PIATETSKY-SHAPIRO; SMYTH (1996).

Na metodologia KDD, é definido primeiramente um objetivo a ser alcançado, e após isso, de um conjunto de base de dados disponível, são selecionados dados considerados importantes para análise. Estes dados escolhidos seguem para uma fase de limpeza e pré-processamento, em que são resolvidos problemas de ruídos, dados inconsistentes e faltantes. Após condicionados, os dados passam por um processo de transformação, em que são modificados e consolidados em um formato apropriado para a mineração, através de operações como sumarização e agregação. A etapa principal do processo KDD é a mineração dados, em que métodos são aplicados para extrair padrões do conjunto de dados. Os padrões determinados passam para uma fase de avaliação, na qual medidas apropriadas são utilizadas para determinar sua eficiência. Como produto final, se obtém o conhecimento, que deve ser apresentado de forma a propor o melhor entendimento possível do conhecimento obtido para o usuário.

Esta dissertação considera como mineração de dados todo o processo de descoberta de conhecimento, como também indicado no trabalho de (ZANGL;

OBERWINKLER, 2004). Para os autores, mineração de dados é o processo de descoberta de novas correlações, tendências e padrões significativos, através da análise de grande conjunto de dados, aplicando para isto técnicas estatísticas, matemáticas e de reconhecimento de padrões. Além disso, os autores apontam sobre a utilidade que ferramentas automatizadas de mineração de dados podem trazer para a atividade de petróleo, já que estas realizam o pré-processamento dos dados brutos, verificação da qualidade deles e extração de informações de uma grande quantidade de dados existentes em um campo de petróleo, promovendo assim melhorias no setor de automatização e otimização da produção.

4.1 Funcionalidades da Mineração de Dados

Técnicas de mineração de dados podem ser aplicadas para diversas funções. Podem ser citadas quatro classes de problemas fundamentais estudados pela atividade de mineração: regras de associação, classificação/regressão, análise de agrupamentos e análise de *outliers*. As diferentes funções se relacionam com os tipos de padrões que serão encontrados nas atividades. Assim, de acordo com os tipos de padrões, as tarefas de mineração podem ser divididas, de forma geral, em duas grandes categorias: descritiva e preditiva. As tarefas descritivas caracterizam as propriedades de um determinado conjunto de dados, e nestas atividades não há variáveis de saída. Nas tarefas preditivas, por sua vez, há um conjunto de variáveis de entrada que irão gerar um modelo que será utilizado para prever uma ou mais variáveis de saída (HAN et al., 2012). A Figura 8 esquematiza como as classes de problemas se mostram divididas, de acordo com os tipos de tarefas de mineração de dados. Esta dissertação concentra suas pesquisas principalmente na área de análise de *outliers*, e atividades preditivas de classificação e regressão.

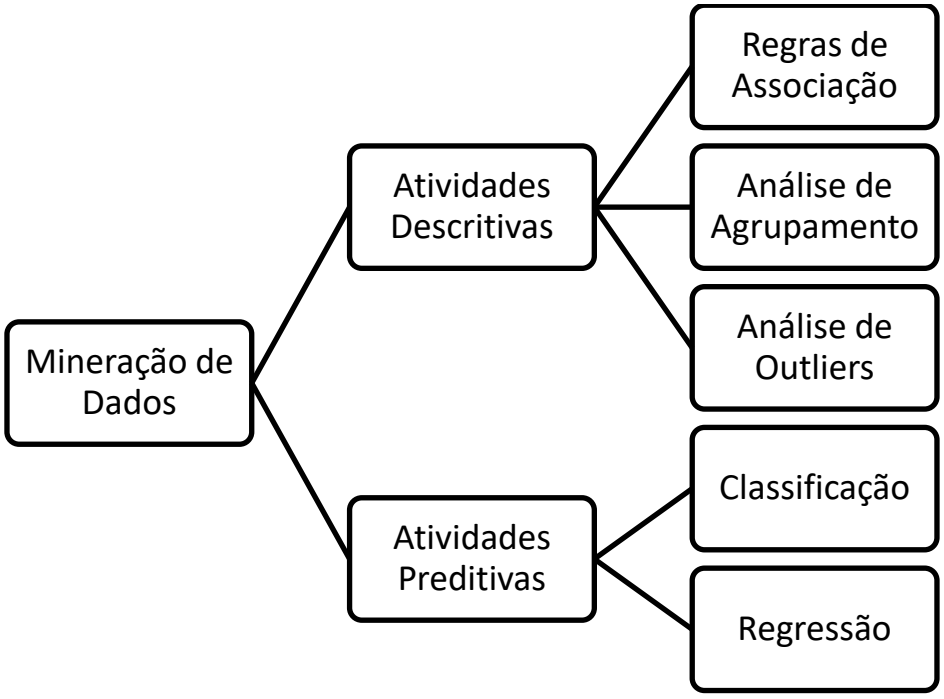


Figura 8: Esquematização das principais funções das atividades de mineração de dados.

4.2 Aprendizado de Máquinas

Mineração de dados é uma área multidisciplinar, que emprega técnicas de diferentes domínios do conhecimento, como: estatística, aprendizado de máquinas, reconhecimento de padrões, áreas de visualização, algoritmos, computação de alta performance e banco de dados. Neste trabalho, são empregadas principalmente as áreas de aprendizado de máquinas e de estatística, além de se ter uma preocupação na elaboração de algoritmos para automatização do processo de análise de dados e visualização desses resultados.

Algoritmos de aprendizado de máquinas foram as principais ferramentas de mineração de dados empregadas nessa dissertação. Aprendizado de máquinas estuda como programas de computadores podem automaticamente aprender a reconhecer padrões complexos e tomar decisões inteligentes, se baseando no conjunto de dados que analisam. De acordo com o tipo de problema que estudam, o aprendizado de máquinas pode ser dividido em: supervisionado, não-supervisionado, semi-supervisionado e ativo (HAN et al., 2012).

Aprendizado supervisionado lida com problemas de classificação e previsão. Nesta tarefa, os dados são divididos em conjunto de treinamento e conjunto de teste. Os dados de treinamento rotulados servem como exemplos para encontrar parâmetros

ótimos que ajustam um determinado modelo. Este modelo deve ser então capaz de prever os rótulos desconhecidos em um conjunto de teste. Se os rótulos são variáveis discretas, é um problema de classificação, se são contínuas, o problema é de regressão. A ideia de aprendizado supervisionado é exatamente se ter no treinamento do modelo os rótulos do problema, e se procurar, a partir do conjunto de dados disponível, as relações existentes entre as variáveis independentes e as variáveis dependentes. Assim, quando novas entradas de variáveis independentes forem inseridas, se quer que o modelo preveja eficientemente o valor da variável alvo.

Aprendizado não supervisionado resolve problemas de agrupamento. Neste processo de aprendizado, os exemplos de entrada não são rotulados. Os métodos podem ser empregados para descobrir classes no conjunto de dados e são geralmente empregados quando não se tem um conhecimento de como são os resultados dos dados.

Neste trabalho são utilizados modelos de aprendizado supervisionado para resolução de problemas de classificação e regressão. Além disso, na análise de identificação de *outliers* foi empregado um modelo do tipo não supervisionado para identificação de dados anômalos.

Os tópicos a seguir abordam, respectivamente, aspectos teóricos dos modelos de classificação e regressão aplicados neste trabalho.

4.2.1 Modelos de Aprendizagem de Classificação

Existem diferentes modelos de classificação na literatura. Foram analisados seis modelos de classificação neste trabalho: Regressão Logística, classificador Naive Bayes, K-Vizinhos mais próximos (do inglês *K-Nearest Neighbor*), Árvore de decisão, Floresta Aleatória (do inglês *Random Forest*) e Máquina de Vetores de Suporte (do inglês, *Support Vector Machine*). Além disso, este item também aborda sobre as principais métricas de erros utilizadas para avaliar o desempenho de modelos de classificação.

4.2.1.1 Regressão Logística

Regressão logística é um modelo preditivo linear em que a variável dependente é qualitativa, enquanto que as variáveis independentes podem ser contínuas ou binárias. Nesta regressão, os erros do modelo não obedecem a uma normal, já que a variável resposta é binária. Além disso, a variância do erro não é constante e sim é uma função

das médias. São considerados então que os erros obedecem a uma distribuição logística, obtendo-se um modelo Logit. Os parâmetros da regressão logística são estimados pelo método da máxima verossimilhança (MONTGOMERY; RUNGER, 2011).

Resumindo, a regressão logística é um caso especial da regressão linear, aplicado para variáveis dependentes categóricas. Caso o modelo de regressão linear fosse utilizado em um caso de respostas discretas 0 e 1, uma reta como a mostrada na Figura 9 seria obtida como resultado. Esta reta não consegue descrever as respostas Y=0 e Y=1. Para contornar o problema, a regressão logística é aplicada na probabilidade de ocorrência de uma das classes, usando a função Logit, que é a curva no formato S, indicada também na Figura 9. É então determinado um valor que define a separação das duas classes, de forma geral é considerado 0,5. Dessa forma, se a resposta da função de regressão logística é maior que 0,5, o Y é classificado como 1, caso contrário, a classe prevista é 0.

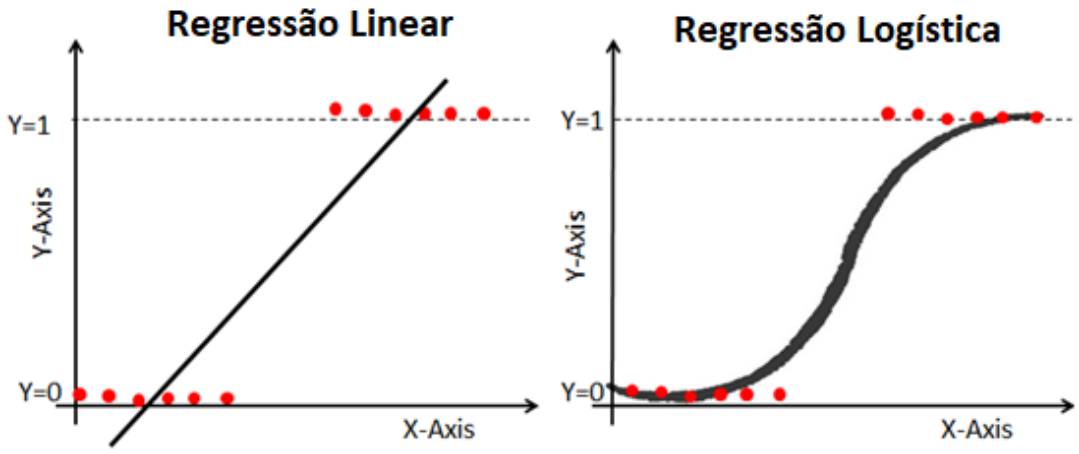


Figura 9: Regressão Linear e Regressão Logística. Fonte: Adaptado de NAVLANI (2018).

4.2.1.2 Classificador Naive Bayes

O classificador Naive Bayes é um classificador probabilístico simples e rápido, geralmente com bom desempenho mesmo para um pequeno conjunto de dados. O método de aprendizado supervisionado tem como fundamentação o teorema de *Bayes*, com a suposição ingênua (*naive*) de que as variáveis são condicionalmente independentes, ou seja, o método desconsidera a correlação entre as variáveis.

Este trabalho aplica o classificador Naive Bayes considerando uma distribuição normal.

4.2.1.3 K-Vizinhos mais próximos (KNN)

K-vizinhos mais próximos, ou KNN é um método não paramétrico, que não utiliza nenhuma função ou modelo específico, e é inteiramente baseado nos elementos da amostra de dados. No método, primeiramente é calculada a distância entre os registros e são identificados os k vizinhos mais próximos de um determinado registro. Após isso, a regra de decisão do vizinho mais próximo é aplicada, na qual uma observação é atribuída ao grupo ao qual a maioria dos seus k-vizinhos mais próximos estão contidos (KIANG, 2003).

A Figura 10 mostra a intuição do método KNN. Considerando as variáveis X1 e X2, e um conjunto de 11 dados, para classificar um novo registro, é necessário primeiramente calcular a distância deste novo ponto aos demais pontos. Após isso, o número de vizinhos é definido. Com a vizinhança delimitada, se verifica qual classe é mais recorrente. Para o caso da Figura 10, o valor definido de k é 3, a classe com maior número de registros na vizinhança é a Classe 2, de forma que o novo registro é rotulado como 2.

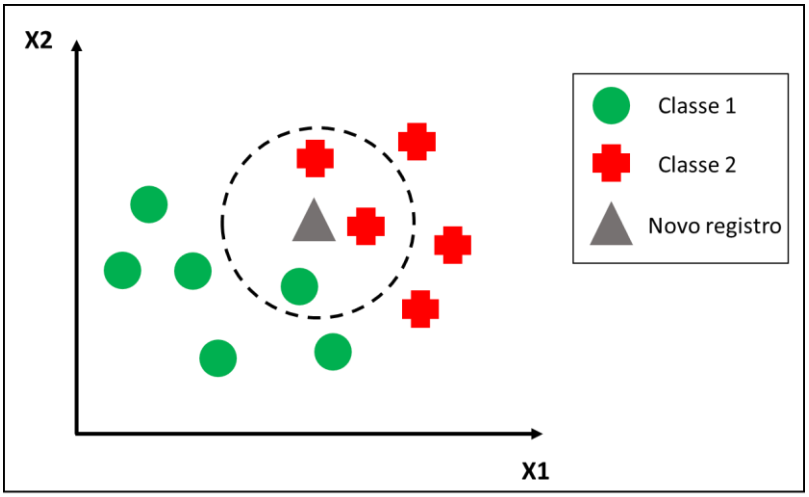


Figura 10: Exemplo do procedimento KNN. Fonte: Elaboração própria.

O desempenho do método KNN é dependente da medida de distância utilizada e da definição do número de vizinhos mais próximos. Dessa forma, é importante calibrar os modelos para encontrar as melhores medidas para o conjunto de dados.

4.2.1.4 Árvores de Decisão

Árvore de decisão é modelada através de um conjunto de decisões hierárquicas sobre as variáveis explicativas, organizadas em estrutura de árvore. Os nós das árvores são os critérios de divisão. Cada nó na árvore representa logicamente um subconjunto do espaço de dados definido pela combinação de critérios de divisão nos nós acima dele (AGGARWAL, 2015). O objetivo principal das árvores de classificação é particionar os dados em grupos menores que sejam homogêneos. A homogeneidade significa que cada nó de divisão contém uma proporção maior de uma determinada classe. Os métodos matemáticos Índice de Gini e entropia são os dois procedimentos mais utilizados para a escolha das partições. Em cada nó, é selecionada a variável que melhor promova a separação das classes, de acordo com o critério utilizado (KUHN; JOHNSON, 2013).

4.2.1.5 Florestas Aleatórias

Florestas Aleatórias (do inglês *Random Forest*) é um conjunto de árvores de decisão, treinadas através do método *Bagging*. O procedimento *Bagging* (*Bootstrapped Aggregation*) combina diferentes modelos de aprendizagem para melhorar a previsão da acurácia do modelo. O método diminui a variância da predição dos modelos, através da amostragem com repetição dos dados originais (AGGARWAL, 2015). Assim, florestas aleatórias constroem e combinam múltiplas árvores de decisão para obtenção de uma melhor acurácia.

No método de florestas aleatórias, a criação de uma árvore é feita selecionando aleatoriamente, através de *bootstrap*, um conjunto de registros com repetição. Além disso, em cada nó de divisão se faz a análise apenas com um subconjunto de variáveis escolhidas aleatoriamente, ao invés de se considerar todas as variáveis. São feitas então diferentes árvores repetindo o procedimento, tendo como resposta um conjunto de modelos de árvores. Durante a classificação de um novo registro, cada árvore irá votar e a classe com maior número de votos será a resposta obtida (HAN et al., 2012).

Florestas aleatórias também são empregadas em problemas de regressão, aplicando o mesmo procedimento descrito. A diferença está na obtenção do resultado final, em que a variável dependente, ou resposta do problema, por ser contínua, será a média dos resultados de todas as árvores.

4.2.1.6 Máquina de Vetores de Suporte

Máquina de Vetores de Suporte (do inglês, *Support Vector Machine*), ou SVM, é o último método de classificação analisado. SVM busca um hiperplano que maximize a margem de separação entre as classes. Para dados linearmente separáveis, o hiperplano de separação é aquele que separa perfeitamente duas classes, e a margem do hiperplano é a soma de sua distância com os pontos de treinamento mais próximos pertencentes a cada uma das classes, chamados vetores de suporte. A Figura 11 exemplifica esses conceitos básicos referente ao método. Além disso, considera-se que a distância do hiperplano de separação até o ponto de treinamento mais próximo de qualquer classe é a mesma. O problema para determinar os coeficientes ótimos do hiperplano é resolvido como uma otimização não linear (AGGARWAL, 2015).

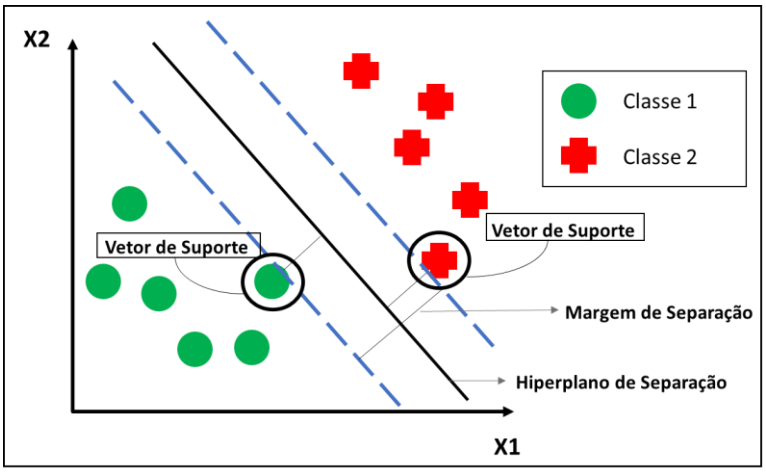


Figura 11: Exemplificação do método SVM. Fonte: Elaboração própria.

Na prática, é atípico encontrar um separador linear perfeito. Entretanto, muitos conjuntos de dados reais podem ser aproximadamente separáveis, casos nos quais a maior parte desses dados estão nos lados corretos dos hiperplanos que caracterizam as margens de separação. Para estes casos em que classes não são perfeitamente linearmente separáveis, as margens podem ser suavizadas, ao se aceitar uma violação das restrições de margem com certa penalidade. O nível de violação de cada restrição de margem para dados de treinamento X_i, em que i é um determinado registro, é dado por uma variável de folga ξ_i. Nos casos em que os dados de treinamento ultrapassam sua margem de separação, essa variável de folga é a distância desses pontos em relação à margem da sua classe. Para pontos que estão corretamente dispostos, o valor de ξ_i é zero. Este comportamento é exemplificado na Figura 12. Neste tipo de problema,

também não é desejável que muitos dados de treinamento tenham valor positivo de ξ_i, de forma que estas violações são penalizadas com um parâmetro de regularização C. Valores pequenos de C resultam em margens relaxadas, enquanto valores grandes de C minimizam erros de dados de treinamento e resultam em margens estreitas (AGGARWAL, 2015).

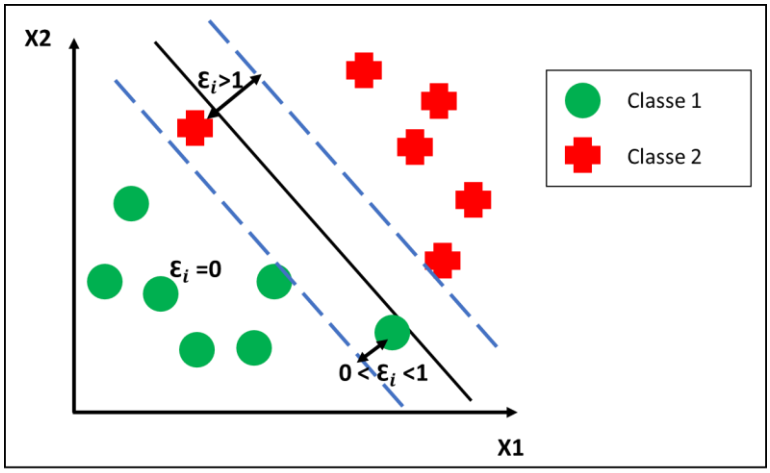


Figura 12: Exemplificação do caso SVM para dados não linearmente separáveis. Fonte: Elaboração própria.

Finalmente, um último aspecto importante é que é possível a resolução de problemas não linearmente separáveis com o método SVM. Para os casos em que nenhum hiperplano linear consegue separar duas classes, o método transforma os dados originais em uma dimensão maior, através de mapeamento não linear, e nesta nova dimensão, procurar por um hiperplano ótimo que melhor separe as classes. Para isto utiliza uma abordagem diferente para aprender os limites de decisão não-lineares, chamada de truque de *kernel*. Esta abordagem é capaz de aprender os limites de decisão sem realizar explicitamente o mapeamento, utilizando funções de núcleo. Esta função núcleo é previamente estabelecida na resolução do problema, e pode ser, por exemplo, linear, de base radial ou polinomial (AGGARWAL, 2015). No caso desta dissertação é utilizado SVM com função núcleo de base radial.

4.2.1.7 Métricas de avaliação para modelos de classificação

Métricas de avaliação são necessárias para verificar o desempenho de um determinado modelo e avaliar a necessidade ou oportunidade de melhorar sua performance. Além disso, ao se estudar um problema, geralmente diferentes modelos

são avaliados, e as medidas de erro servem como parâmetros para definir qual melhor modelo para o caso estudado.

Para problemas de classificação, é necessário que o conjunto de dados seja dividido em uma série de treinamento do modelo, e uma série de teste para avaliar sua performance. Assim, um determinado modelo é ajustado utilizando a série de treino e esse modelo é usado para prever resultados utilizando os dados de entrada da série de teste. Após isso, é possível avaliar os resultados, comparando os resultados obtidos pelo modelo com os resultados reais, contidos na série de teste. Diferentes métricas podem ser utilizadas para avaliar e comparar o desempenho dos modelos. A Tabela 3 mostra as medidas usuais para avaliar os modelos de classificação (HAN et al., 2012).

Tabela 3: Medidas de Avaliação para Modelos de Classificação.

Medidas de Avaliação	**Formulação**
Acurácia	$\frac{VP + VN}{P + N}$
Erro	$\frac{FP + FN}{P + N}$
Sensibilidade	$\frac{VP}{P}$
Especificidade	$\frac{VN}{N}$
Precisão	$\frac{VP}{VP + FP}$
Medida F	$\frac{2 * Precisão * Recuperação}{Precisão + Recuperação}$

Sendo VP, VN, FP, FN, P, N respectivamente o número de verdadeiros positivos, verdadeiros negativos, falsos positivos, falsos negativos, total de positivos reais e total de negativos reais.

VP e VN são a quantidade de valores que o modelo categorizou corretamente. Ou seja, representam o número de acertos do modelo, o modelo definiu a classe positiva e os dados reais eram positivos, assim como definiu a classe negativa, e realmente os dados reais eram negativos. FP e FN representam a quantidade de valores categorizados errados pelo modelo. FP é o número de valores que o modelo categorizou como

positivo, mas na realidade era negativo, e FN é o contrário, o modelo colocou na classe negativa, mas na realidade era positivo.

Os termos podem ser sumarizados através da matriz de confusão, mostrada na Tabela 4. Na matriz, as linhas são as quantidades de dados reais específicos de cada classe, e as colunas são as quantidades de elementos que o classificador previu em cada classe. Dessa forma, elementos da diagonal principal (VP e VN) correspondem a quantidade prevista corretamente pelo modelo para cada classe, e os demais termos representam a quantidade de termos classificadas de forma incorreta. O modelo tem uma boa acurácia quando maior parte dos elementos estiverem na diagonal principal (HAN et al., 2012).

Tabela 4: Matriz de confusão para classificação de classes.

	CLASSES PREVISTAS			
CLASSES REAIS		**SIM**	**NÃO**	**TOTAL**
	SIM	VP	FN	P
	NÃO	FP	VN	N
	TOTAL	P’	N’	P + N

Observando a Tabela 4, é possível obter um melhor entendimento dos conceitos das métricas mostradas na Tabela 3. Acurácia e erro são medidas globais, que indicam respectivamente a proporção de acertos e erros do modelo. Essas medidas, entretanto, não são adequadas para avaliar casos de problemas desbalanceados, ou seja, problemas nos quais a quantidade de registros de uma das classes é muito superior à das demais. Por exemplo, considerando um caso em que 90% dos dados pertençam a uma determinada classe, obter uma acurácia de 90% não significa que o modelo apresentou bom desempenho, se ele conseguir acertar apenas a classe com maior número de registros. Conjunto de dados desbalanceados é também o caso do problema estudado neste trabalho, em que a quantidade de testes de produção válidos é muito superior à de testes de produção inválidos.

As métricas de sensibilidade e especificidade, mostradas na Tabela 3, são mais adequadas para a avaliação de séries desbalanceadas, pois analisam cada classe individualmente. Sensibilidade é a proporção de verdadeiros positivos, ou seja, dentre os elementos da classe positiva, quais realmente foram classificados como positivos,

enquanto que especificidade é a taxa de verdadeiros negativos, ou seja, elementos na classe negativa que foram corretamente classificados como negativos.

Outra métrica muito importante para análise de séries desbalanceadas é a curva ROC (*Receiver Operating Characteristic*). Para um determinado modelo, a curva mostra um *tradeoff* entre a taxa de verdadeiros positivos (sensibilidade) e a taxa de falsos positivos (1 – especificidade). Para um conjunto de dados de teste e um modelo, a taxa de verdadeiros positivos (TVP) é a proporção de positivos que foram corretamente classificados pelo modelo, ou de acordo com a Tabela 4, TVP = VP/P. Da mesma forma, a taxa de falsos positivos (TFP) é a proporção de negativos que foram classificados como positivos, ou seja, TFP = FP/N. A curva marca essas duas taxas em diferentes limiares de classificação. Para todo conjunto de teste, são testados diferentes probabilidades ou pesos para determinada classe, de forma que a curva ROC fornece um resumo de todos os resultados possíveis da matriz de confusão (HAN et al., 2012).

Para sintetizar o resultado obtido pela curva ROC, pode-se utilizar a métrica AUC (*Area Under the ROC Curve*), que é a área abaixo da curva ROC. A medida AUC resume a informação da curva. AUC varia de 0 até 1. Valores de AUC iguais a zero indica que as previsões estão totalmente erradas, enquanto AUC igual a 1 mostra que o modelo está perfeitamente acurado. Um valor de AUC igual a 0,5, que é o limiar de divisão das classes, indica que o modelo não é capaz de diferenciar entre a classe positiva e negativa. Um bom modelo de classificação deve estar o mais próximo possível de 1 (HAN et al., 2012).

4.2.2 Modelos de Aprendizagem de Regressão

Assim como nos modelos de classificação, existem diferentes modelos de regressão na literatura. Neste trabalho foram analisados quatro modelos de regressão: regressão linear múltipla, regressão por vetores de suporte (do inglês, *Support Vector Regression*), árvore de regressão e floresta aleatória (do inglês *Random Forest*) para regressão. Como o procedimento de florestas aleatórias é semelhante para problemas de classificação e regressão, este método não será novamente exposto. Este item ainda aborda as métricas de erros utilizadas para avaliar o desempenho de modelos de regressão.

4.2.2.1 Regressão Linear Múltipla

Regressão linear múltipla é uma generalização da regressão linear simples, para os casos que o modelo de regressão contém mais de uma variável explicativa. Assim, uma variável dependente ou resposta do modelo Y é relacionada com k variáveis independentes. O modelo de regressão linear múltipla é dado pela Equação 4.

$$Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_k x_k + \varepsilon \qquad \text{Equação 4}$$

Em que o número de variáveis independentes é igual a k, e os parâmetros β_j são os coeficientes da regressão, para j = 0, 1,...k. O modelo descreve um hiperplano em um espaço de dimensão k. Os coeficientes da regressão linear múltipla, assim como na simples, podem ser estimados pelo método dos mínimos quadrados (MONTGOMERY; RUNGER, 2011).

4.2.2.2 Regressão por Vetores de Suporte

O conceito de máquinas de vetores de suporte (SVM) para problemas de classificação, descrito no item 4.2.1.6, pode ser estendido para problemas de regressão. Esta técnica é considerada robusta, pois busca minimizar o efeito de *outliers* sobre as equações de regressão. Outro aspecto importante é que existem diferentes técnicas de regressão de vetores de suporte, sendo a usual e considerada neste trabalho é a do tipo ε-insensível. Regressão por Vetores de Suporte (SVR) utiliza uma função similar com a função de Huber (HUBER, 1964). Nesta, são utilizados resíduos quadrados para resíduos de pequena escala e resíduos absolutos para aqueles de grande escala, com o objetivo de minimizar o efeito dos métodos dos mínimos quadrados que considera com mesmo peso observações distantes e próximas da tendência geral dos dados, o que pode comprometer com a eficiência do modelo (KUHN; JOHNSON, 2013).

Na regressão SVR, um limite de tolerância é definido pelo usuário (o parâmetro ε). Os dados com resíduos dentro deste limite não contribuem para o ajuste da regressão, enquanto os pontos com desvio absoluto superior a esse limite têm uma contribuição de escala linear. Como a escala é linear e não quadrática, os pontos muito destoantes não têm um impacto muito grande no ajuste do modelo. Figura 13 mostra uma exemplificação da metodologia SVR para o caso linear. Na figura, o limite ε cria

um tubo em torno da função de regressão. Somente os pontos fora deste limite serão considerados no ajuste do modelo (KUHN; JOHNSON, 2013).

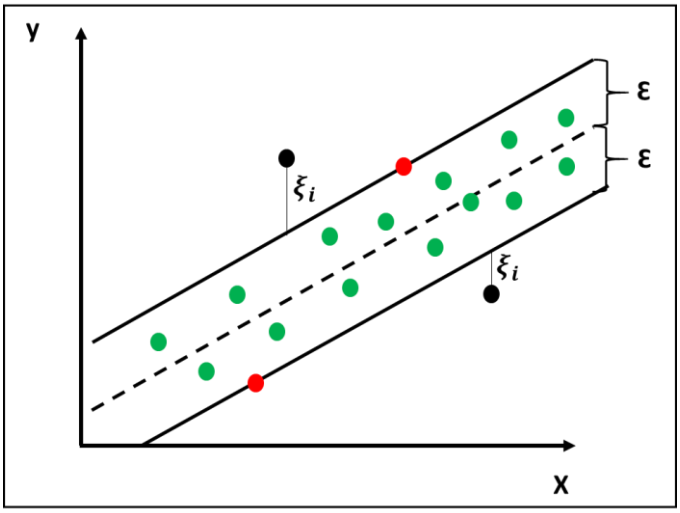


Figura 13: Exemplificação do modelo SVR ε-insensível. Fonte: Elaboração própria.

Assim como no caso SVM, o modelo de regressão também aplica funções núcleo para os casos não lineares. Neste trabalho, o SVR também utiliza função núcleo de base radial.

4.2.2.3 Árvores de Regressão

O método de árvore de decisão pode ser aplicado tanto para classificação como para a regressão. A diferença é que para o caso da regressão, a resposta do modelo, ou seja, a variável dependente, é contínua. Para construir a árvore de regressão é necessário determinar o preditor e seu respectivo valor para divisão, a profundidade da árvore e a equação de predição em cada nó. Na metodologia mais frequentemente utilizada para construção da árvore, o modelo inicia com todo conjunto de dados S, e procura todos os valores de cada variável preditora (independente) para encontrar a variável e o valor da divisão que particiona os dados em dois grupos (S_1 e S_2), de forma a minimizar a soma total dos quadrados dos erros (SSE), mostrada na Equação 5 (KUHN; JOHNSON, 2013).

$$SSE = \sum_{i \in S_1} (y_i - \bar{y}_1)^2 + \sum_{i \in S_2} (y_i - \bar{y}_2)^2 \qquad \text{Equação 5}$$

De acordo com a Equação 5, $\bar{y}_1$ e $\bar{y}_2$ são as médias das séries de dados dos grupos S_1 e S_2, respectivamente.

4.2.2.4 Métricas de avaliação de modelos de regressão

Existem diferentes métricas para avaliação dos modelos de regressão. Neste trabalho se adotou a medida de erro Raiz do Erro Quadrático Médio (do inglês, *Root Mean Squared Error*, RMSE) e o coeficiente de determinação ajustado (R^2_{aj}). Equação 6 e Equação 8 fornecem a formulação de RMSE e R^2_{aj}, respectivamente. O cálculo de R^2_{aj} exige a determinação do coeficiente de determinação, mostrado na Equação 7.

$$\text{RMSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2} \qquad \text{Equação 6}$$

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2} \qquad \text{Equação 7}$$

$$R^2_{aj} = 1 - (1 - R^2)\frac{n-1}{n-p-1} \qquad \text{Equação 8}$$

Em que n é o número de amostras, y_i é o valor real da variável dependente e $\hat{y}_i$ é o valor ajustado, para i= 0, 1,...,n; $\bar{y}$ é o valor médio da variável e p é o número de variáveis independentes.

RMSE pode ser interpretado como a distância média entre os valores observados e os valores previstos. E o coeficiente de determinação R^2 indica a proporção de informação dos dados explicada pelo modelo de regressão (KUHN; JOHNSON, 2013). Entretanto, em função da forma da equação de R^2, percebe-se que a adição de uma variável independente (variáveis de entrada) ao modelo de regressão sempre faz com que o coeficiente de determinação melhore, independente se a variável adicionada trouxe melhorias ao ajuste do modelo. Dessa forma, para modelos com mais de uma variável independente (variáveis de entrada), o caso desse trabalho, é necessário estabelecer um fator de penalização para adição de variáveis que não melhorem o

modelo. Por isso se utiliza como métrica o coeficiente de determinação ajustado (R_{aj}^2), que leva em consideração os graus de liberdade do modelo.

4.3 Trabalhos Relacionados a Mineração de Dados na Atividade de Produção de Petróleo

Após a abordagem dos conceitos básicos da área de análise de dados, este último tópico procura mostrar alguns trabalhos da literatura que relatam o uso de métodos de mineração de dados e de novas tecnologias inteligentes aplicadas ao processo de produção de petróleo.

ZANGL e OBERWINKLER (2004) mostram como modelos preditivos de mineração de dados podem ser aplicados para acelerar e otimizar a produção de reservatórios. O artigo apresenta uma ferramenta criada em que é aplicado um tipo especial de redes neurais, os mapas auto-organizáveis (do inglês, *self organizing map*, SOM). O modelo criado pode ser utilizado tanto para controle de qualidade de dados de produção de poços como para determinar o comportamento de cada poço. No primeiro caso, o modelo correlaciona as relações entre todos os parâmetros de produção de um determinado poço. No exemplo mostrado, 10 parâmetros adquiridos diariamente são usados como dados de entrada para o modelo de redes neurais: tamanho do *choke*, pressão do revestimento, razão gás-óleo, vazão bruta, produção de óleo, fração de água, pressão e temperatura do separador, pressão e temperatura da cabeça do poço. Assim, a partir da aquisição de novos dados no sistema, a rede neural é acionada e calcula os valores de todos os parâmetros de entrada, usando o modelo aprendido com os dados históricos. Após isso, os valores calculados de cada parâmetro são comparados com os valores adquiro, permitindo uma avaliação e detecção de valores anômalos.

Além de ser utilizado para controle de qualidade, descobrindo relações entre todo o conjunto de dados de entrada, o artigo ainda emprega o modelo SOM para determinar o comportamento do poço. SOM classifica todo conjunto de dados de entrada, de acordo com sua similaridade e cria *clusters* automaticamente. Neste caso, o modelo pode identificar todos os diferentes comportamentos de produção que o poço passou ao longo da sua vida produtiva, como por exemplo, períodos de fechamento, períodos de baixo desempenho em função de ajustes ineficientes de métodos de elevação artificial e períodos de alta performance. Cada período que o poço passou está

associado a um determinado grupo. Assim, quando um novo ponto é adquirido, o modelo calcula, além dos valores esperados de cada parâmetro, o valor do *cluster* no qual está associado, podendo se ter a representação da situação do poço. Além do método apresentado, o artigo ainda reforça sobre a importância de automatização da aquisição e manipulação de dados para um melhor controle do sistema da produção.

Do projeto piloto de operações de campos de óleo digitais inteligentes (iDOF) que está sendo conduzido no campo Sabriya, descrito no item 3.3, um dos módulos desenvolvidos é sobre bombas. O artigo de VELASQUEZ et al. (2013) descreve o módulo BCS (bombas centrífugas submersas) que integra uma das partes do projeto iDOF. No campo em estudo, bombas centrífugas submersas (BCS) são as mais utilizadas para a produção. O uso de bombas, entretanto, está associado a certos problemas que podem ocasionalmente surgir, como desgaste, entupimento na entrada do equipamento, interferência de gás, vazamento na tubulação e problemas de viscosidade dos fluidos. Assim, o foco principal do módulo é monitorar o desempenho das bombas em tempo real, para permitir que os operadores detectem e resolvam os cinco problemas apresentados de forma mais rápida.

O procedimento é conduzido através de um fluxo de trabalho que combina métodos analíticos para avaliação dos parâmetros operacionais do sistema, e um sistema que utiliza lógica *Fuzzy* para identificação de mau funcionamento do sistema. Neste, o comportamento dos parâmetros na presença dos problemas citados é mapeado, e a probabilidade de ocorrência, em tempo real, de cada um dos cinco problemas é obtida como resultado final e indicada para os operadores.

Ainda na linha de tecnologias inteligentes para detecção de falhas em bombas centrífugas submersas, ABDELAZIZ et al (2017) também avaliam falhas relacionadas à BCS, mas utilizando uma abordagem diferente. Os autores aplicam Análise de Componentes Principais (PCA) para detectar possíveis falhas durante a utilização de BCS assim como prever o tempo de operação restante antes da ocorrência da falha. Com o uso de PCA, *clusters* foram gerados com dados de histórico das bombas, possibilitando a identificação de regiões estáveis e não estáveis para diagnóstico e prognóstico da BCS. Foram utilizados históricos de cinco instalações BCS e o estudo concluiu que PCA pode ser empregado como ferramenta para identificação de mudanças dinâmicas no sistema BCS e detecção de problemas.

Em relação a técnicas de aprendizado para classificação de testes, SUBRAHMANYA et al. (2014) revisam métodos de aprendizado semi-supervisionado e ativo e a eficiência destes é avaliada em um exemplo prático da indústria de petróleo. Esses procedimentos são aplicados quando, dado um conjunto de registros disponíveis, apenas uma pequena parte deles já está classificada, como, por exemplo, teste válido ou não válido. O uso desses métodos permite reconhecer padrões e determinar a validade dos demais. O artigo então cria um modelo estatístico, utilizando tais técnicas, para determinar a validade dos testes de medição de vazões em um separador de teste. Muitos dos testes podem não ser aceitáveis devido a uma série de razões. O objetivo do artigo é reconhecer padrões de forma a capturar e reconhecer características importantes de um teste classificado como bom e ruim.

Para realizar o estudo, primeiramente, o artigo propõe rearranjar a matriz de similaridade dos registros para uma melhor visualização da estrutura dos dados. A partir dos *clusters* formados nessa reorganização, são selecionados os primeiros pontos para serem rotulados. Em uma segunda etapa, utiliza-se os rótulos disponíveis e os dados não rotulados para aprender a rotulação de todo o conjunto de dados utilizando aprendizado semi-supervisionado. Na terceira etapa, distribuições de probabilidade previstas são usadas na associação de classes para identificar os pontos de dados com maior incerteza e apresentá-los ao usuário para serem rotulados pelos especialistas. As duas etapas anteriores são repetidas até que a alteração nos rótulos previstos esteja abaixo de um pequeno limite. Dentre as análises conduzidas, percebeu-se que os melhores resultados são obtidos ao se utilizar aprendizado ativo para selecionar os pontos a serem rotulados e aprendizado semi-supervisionado para propagar os rótulos aprendidos para pontos de dados não rotulados.

Detecção de falhas pode ser feita utilizando métodos de aprendizado de máquinas para identificação de *outliers* no sistema. CHAUDHARY e LEE (2016) desenvolveram um método robusto não supervisionado para identificar *outliers* em dados de vazão e pressão usados para curva de declínio e previsão, análises de pressão e vazão transiente, e fluxo de trabalhos semelhantes. A metodologia de identificação de *outliers* desenvolvida é baseada no método LOF (*Local Outlier Factor*) (BREUNIG et al., 2000) que leva em consideração a densidade local ao redor de um dado, e a localização é quantificada em função da distância do ponto até seus k vizinhos mais próximos. É um método não paramétrico, sem necessitar assumir um modelo correto a

priori para modelar os dados. O método proposto é validado através de exemplos sintéticos gerados usando modelos numéricos de poços fraturados hidraulicamente em múltiplos estágios em reservatórios não convencionais.

Um tópico muito importante nesse trabalho é a utilização de técnicas avançadas de mineração de dados para previsão das variáveis do processo, como vazões e pressões. Nesta linha de pesquisa, pode ser citado o artigo de (CAO et al., 2016) que utilizam redes neurais artificiais, um algoritmo de Aprendizado de Máquinas, para prever a produção de poços em dois cenários: previsão em poços existentes e previsão em um novo poço que ainda será perfurado. A técnica é aplicada em um reservatório não convencional, utilizando como informações de entrada, dados de mapas geológicos, histórico de produção, e restrições operacionais, como pressão na cabeça do poço. O método foi comparado com técnicas convencionais de curvas de declínio, os métodos Arps (1945), Duong (2010), SEPD (VALKO; LEE, 2010) e lei da Potência (ILK et al., 2010). A pressão real da cabeça do poço foi utilizada como entrada do modelo de redes neurais, e os resultados de validação mostraram que o modelo foi o que obteve resultados de previsão mais próximos dos dados reais. Além disso, o artigo mostra que métodos baseados em dados podem ser aplicados como um intermédio entre um modelo robusto e técnicas rápidas para avaliação do comportamento de poços. Além de fornecer suporte a técnicas analíticas.

BALAJI et al. (2018) fazem uma revisão das técnicas baseadas em dados que são aplicadas em diferentes ramos da indústria de óleo e gás. Os autores mostram as vantagens e desvantagens de cada uma destas metodologias, e exemplos do emprego destas técnicas na indústria. Os modelos baseados em dados são utilizados em diferentes áreas, como nas áreas de gerenciamento e simulação de reservatório, otimização da produção e perfuração, automatização de perfuração em tempo real e manutenção das instalações. Nas aplicações na etapa de produção, pode-se citar, por exemplo, técnicas de SVM para classificar as condições que, durante a produção de petróleo, levam ao início da produção de areia; SVM aplicado para previsão da temperatura de formação de hidrato para diferentes tipos de gases; e rede neural para controle e automação da produção de gás de xisto. O artigo ainda ressalta sobre o cuidado que se tem que ter para a utilização cega de modelos baseados somente em dados. Em situações em que não se tem quantidade suficiente de dados ou que o sistema não esteja estável durante o período coberto pelo modelo, o algoritmo pode não modelar corretamente. O artigo

indica que existe uma tendência atual de se usar modelos híbridos que combine diferentes métodos baseados em dados e físicos para gerar uma solução para o problema.

KHAN et al. (2019) empregam algoritmos de aprendizado de máquinas para desenvolver uma correlação que consiga prever a vazão de óleo em poços que utilizam *gas lift* como método de elevação artificial. O objetivo do artigo é fornecer soluções simples que possam ser universalmente aplicáveis. Foram utilizados os métodos Sistema de Inferência Adaptativo Neuro-Difuso (ou ANFIS – Adaptative Neural Fuzzy Inference System) e SVM. O conjunto de dados utilizado era compostos de 1500 registros de separadores de teste, com os parâmetros tamanho do choke, pressão no topo, temperatura no fundo e no topo, e grau API do óleo como dados de entrada, e a vazão de óleo medida nos separadores como variável de saída. O modelo empírico proposto visa estimar a vazão de óleo para qualquer conjunto de dados em que os parâmetros de entrada estejam dentro do intervalo do modelo. O modelo desenvolvido utilizando redes neurais foi validado testando contra as correlações empíricas normalmente utilizadas, e consegue prever a vazão com 99% de acurácia. Um dos empregos do método desenvolvido é na validação de testes de produção.

A seguir é mostrado um resumo com os trabalhos expostos anteriormente neste item.

Tabela 5: Revisão bibliográfica de trabalhos de técnicas de dados na atividade de produção de petróleo

Trabalhos	Validação de Testes	Produção	Método
ZANGL e OBERWINKLER (2004)		✓	Mapas auto-organizáveis (SOM)
AL-ABBASI et al. (2013)		✓	Redes neurais, lógica Fuzzy e RCA.
VELASQUEZ et al. (2013)		✓	Lógica Fuzzy para identificação de mau funcionamento da BCS.
ABDELAZIZ et al (2017)		✓	Análise de Componentes Principais (PCA) para detectar possíveis falhas na BCS.
SUBRAHMANYA et al. (2014)	✓	✓	Aprendizado semi-supervisionado para classificação de testes
CHAUDHARY e LEE (2016)		✓	LOF (Local Outlier Factor) para detectar falhas.
CAO et al. (2016)		✓	Redes neurais artificiais para prever vazão.
BALAJI et al. (2018)		✓	Revisão de métodos, como SVM, redes neurais.
KHAN et al. (2019)	✓	✓	SVM e redes neurais para previsão da vazão de óleo.

5 PROCEDIMENTO METODOLÓGICO

Este capítulo aborda o procedimento metodológico utilizado para o desenvolvimento de uma ferramenta automatizada de validação de testes de produção de petróleo, que determine, durante a realização do novo teste, se este é válido ou não, além de fornecer um intervalo de valores esperados para as vazões de óleo, água e gás, para que durante a realização do teste se possa verificar se as vazões estão de acordo com seu comportamento esperado. Para isso, o trabalho pode ser dividido em quatro etapas principais, conforme mostrado na Figura 14.

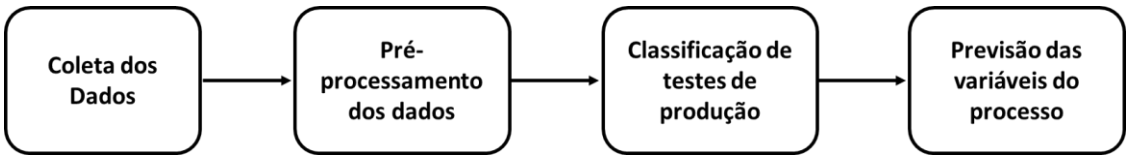


Figura 14: Fluxo esquemático da metodologia do trabalho.

Em uma etapa inicial, para cada poço, boletins de testes de produção são coletados e condensados em uma tabela de histórico de testes. Estes dados passam então para uma etapa de pré-processamento para identificação de dados anômalos. Após isso, estes dados tratados entram na fase de classificação de testes de produção, que tem como objetivo classificar um teste em válido e inválido, de acordo com as variáveis obtidas. Finalmente, a última etapa é a previsão das variáveis do processo, em que as vazões de óleo, água e gás são previstas para o novo teste. Nesta fase, para cada variável, é obtido um valor ajustado e um intervalo de previsão esperado. A partir dos resultados obtidos nas etapas anteriores, pode-se verificar durante a realização do teste se os valores obtidos estão dentro do intervalo esperado e se o teste pode ser classificado em válido ou inválido. Cada uma das etapas da metodologia é descrita com mais detalhes nos tópicos a seguir.

5.1 Coleta de Dados

Em uma primeira etapa, boletins de testes de produção de poços de petróleo, válidos e não válidos, obtidos durante um período de produção são coletados. Para cada poço, é necessário compactar as informações dos boletins em uma única tabela, em que cada coluna corresponde a uma variável e cada linha caracteriza as variáveis de um

determinado teste, juntamente com uma categoria que o define como válido ou não válido.

Dentre as variáveis selecionadas para formular a tabela, estão aquelas de medição direta, cujas medidas são computadas por hora durante a realização do teste e o resultado final é a média desses registros. Essas variáveis são mostradas na Tabela 6.

Tabela 6: Variáveis de obtenção direta dos boletins de teste de produção.

VARIÁVEL	**DESCRIÇÃO**	**UNIDADE**
P1	Pressão no fundo do poço	Kgf/cm²
P2	Pressão na cabeça do poço	Kgf/cm²
P3	Pressão de chegada na plataforma	Kgf/cm²
P4	Pressão no separador de teste	Kgf/cm²
T1	Temperatura no fundo do poço	ºC
T2	Temperatura na cabeça do poço	ºC
T3	Temperatura de chegada	ºC
T4	Temperatura no separador de teste	ºC
Qbruta	Vazão de líquido produzido	m³/d
Qágua	Vazão de água produzida	m³/d
Qóleo	Vazão de óleo produzida	m³/d
Qgt	Vazão total de gás	m³/d
Qgp	Vazão de gás produzido	m³/d
Qgl	Vazão de *gas lift*	m³/d
Pgl	Pressão a jusante de *gas lift*	Kgf/cm²

Além disso, os boletins de testes de produção contêm variáveis de obtenção indireta, a partir das medidas obtidas pela Tabela 6 ou por outros testes, e outras informações características do teste, como a data e a duração do teste. Essas medidas dos boletins também foram coletadas para formular a tabela de dados e estão indicadas na Tabela 7.

Tabela 7: Variáveis de obtenção indireta dos boletins de teste de produção.

VARIÁVEL	DESCRIÇÃO	UNIDADE
Data	Data que teste foi realizado	
t	Tempo de duração do teste	horas
FE	Fator de Encolhimento	
RS	Razão de solubilidade	m^3/m^3
RGO	Razão gás-óleo	m^3/m^3
RGLI	Razão gás-líquido	m^3/m^3
BSW	Fração de água produzida	%

Finalmente, a partir de registros externos, é possível obter informações sobre características do reservatório e do poço no período que o teste está sendo realizado, assim como a identificação se aquele teste foi considerado válido ou não. Essas últimas informações que compõem a tabela de dados estão indicadas na Tabela 8.

Tabela 8: Informações complementares.

VARIÁVEL	DESCRIÇÃO	UNIDADE
IP	Índice de Produtividade	$(m^3/d)/(Kgf/cm^2)$
Pe	Pressão estática do reservatório	Kgf/cm^2
Status	Teste válido ou inválido	

Assim, as tabelas dos dados são constituídas pelas variáveis mostradas na Tabela 6, Tabela 7 e Tabela 8. Estas variáveis serão utilizadas nas etapas posteriores. Uma rotina foi implementada em *Python* para acessar essas informações dos boletins de testes de produção e planilhas complementares. Dessa forma, para cada poço de petróleo analisado, a rotina acessava as variáveis de cada teste e completava a tabela de dados. Além disso, a partir da variável Data, mostrada na Tabela 7, foi gerada uma outra categoria, Dias, que indica o período decorrido de testes, para facilitar as operações e a consideração do tempo na análise. O comportamento das variáveis obtidas durante os testes de produção pode variar com o tempo de produção, de forma que é importante fazer a análise considerando uma possível influência do fator dia.

5.2 Pré-Processamento de Dados

Após a fase de coleta de dados, as tabelas criadas passam para fase de pré-processamento em que são identificados e removidos dados inconsistentes. Os dados coletados dos boletins de testes de produção podem conter dados anômalos, decorrentes de diferentes causas, como erros de digitação e falhas nos equipamentos de monitoramento, que se não forem corrigidos, podem levar a análises incorretas e dificuldades no processo de modelagem. Dessa forma, é necessário tratar o conjunto de dados, aplicando métodos para identificar e remover essas informações inconsistentes, denominadas *outliers*.

Em uma análise prévia, determinados problemas foram verificados e corrigidos sem aplicação de métodos. Em alguns casos, aparelhos de medição de pressões e temperaturas não funcionam, e as medições são registradas com o valor nulo. Entretanto, é impossível fisicamente, para os poços analisados, ter medidas de pressão e temperatura nulas. Então, inicialmente são removidos dessas variáveis todos os valores nulos. Além disso, as vazões são influenciadas pela diferença de pressões ao longo do sistema de produção, tendo sua maior pressão no fundo do poço e a menor pressão ao chegar na plataforma. Assim, para facilitar a identificação de dados anômalos levando em consideração esse fenômeno físico, foram geradas mais duas variáveis na tabela analisada, $\Delta P1$ e $\Delta P2$, conforme mostradas nas Equação 9 e Equação 10, respectivamente. Os casos em que valores negativos foram encontrados para essas variações de pressão são indicativos de que a medida de pressão está incorreta.

$$\Delta P1 = P1 - P2 \qquad \text{Equação 9}$$

$$\Delta P2 = P2 - P3 \qquad \text{Equação 10}$$

Em que $\Delta P1$ representa a diferença de pressão do fundo do poço até a cabeça do poço, e $\Delta P2$ a diferença de pressão da cabeça do poço até a chegada da plataforma.

Após feitas essas modificações, foram aplicados três métodos de identificação de *outliers*. Quanto ao número de variáveis analisadas, os métodos podem ser classificados como univariados e multivariados. Além disso, em relação a utilização de modelos estatísticos, os métodos de análise de *outliers* podem ser paramétricos, em que

utilizam modelos de distribuição dos dados em sua formulação, e não paramétricos, que são os métodos livres de modelos e baseados em técnicas de proximidade (BEN-GAL, 2005).

Z-score modificado (IGLEWICZ, B. AND HOAGLIN, 1993) foi o primeiro método aplicado. É um método univariado e paramétrico, baseado na mediana e desvio absoluto da mediana da série. Considerando uma série temporal em que o eixo das ordenadas é a variável em análise e o das abscissas é o tempo decorrido, é possível aplicar o método Z-score modificado. No método, para cada amostra, um parâmetro chamado M_i é calculado usando a Equação 11.

$$M_i = \frac{0{,}6745}{MAD}\ (x_i - \tilde{x}) \qquad \text{Equação 11}$$

Na Equação 11, x_i é o valor da amostra i, $\tilde{x}$ é a mediana das amostras, e MAD é o desvio absoluto da mediana, dado pela mediana$\{|x_i - \tilde{x}|\}$. Para calcular o MAD, se calcula a diferença entre cada ponto e a mediana das amostras, e a medida MAD será dada pela mediana destas diferenças calculadas. O valor da constante 0,6745 da Equação 11 é em função de E(MAD) = 0,6745σ, para um valor de amostras elevado. Usando o método de Z-score modificado, *outliers* são identificados quando $|M_i| > D$, e geralmente se considera D = 3. A vantagem do método é que ele utiliza como medidas a mediana e o MAD que são medidas mais robustas de tendência central e dispersão, respectivamente.

Para identificar *outliers* usando o método de Z-score modificado, é necessário, primeiramente, ajustar os dados da variável analisada com um modelo de regressão adequado que consiga identificar a tendência da variável analisada ao longo do tempo. Como muitas variáveis modificam suas tendências ao longo do tempo, é importante considerar esta análise temporal. Assim, no modelo de regressão ajustado, o eixo das abcissas representa os dias do histórico de produção analisado e o eixo das ordenadas é a variável sob análise. Depois, os desvios são calculados pela diferença entre dados reais e os valores ajustados para remover a temporalidade da série. Z-score modificado é então aplicado a esses desvios e os *outliers* são identificados. O modelo SVR (regressão por vetores suporte, do inglês, *Support Vector Regression*) (VAPNIK, 1995), conforme mostrado no item 4.2.2, foi o método de regressão utilizado para modelar o conjunto de dados de cada variável analisada.

Um dos problemas enfrentados pelo método Z-score modificado é que a eficiência do método é dependente da qualidade do ajuste do modelo de regressão. Além disso, o procedimento não pode ser aplicado em análises multidimensionais, que também necessitam ser estudadas neste trabalho.

O segundo método aplicado é baseado em distâncias entre registros. Estes métodos conseguem analisar todos os registros e variáveis simultaneamente, e não consideram o modelo de distribuição dos dados. Neste trabalho, a métrica utilizada foi a distância Euclidiana (HAN et al., 2012). Considerando que i = $(x_{i1}, x_{i2}, \ldots, x_{ip})$ e j = $(x_{j1}, x_{j2}, \ldots, x_{jp})$ dois objetos descritos por p variáveis numéricas, a fórmula da distância Euclidiana é dada pela Equação 12.

$$d_{i,j} = \sqrt{\left(x_{i1} - x_{j1}\right)^2 + \left(x_{i2} - x_{j2}\right)^2 + \ldots + \left(x_{ip} - x_{jp}\right)^2} \qquad \text{Equação 12}$$

Em que $d_{i,j}$ é a distância entre os registros i e j, em uma série com p variáveis.

Após o cálculo das distâncias, a matriz de distâncias entre registros D é obtida, conforme exemplificada na Figura 15. Nesta, os pontos com cores mais forte são potenciais *outliers*.

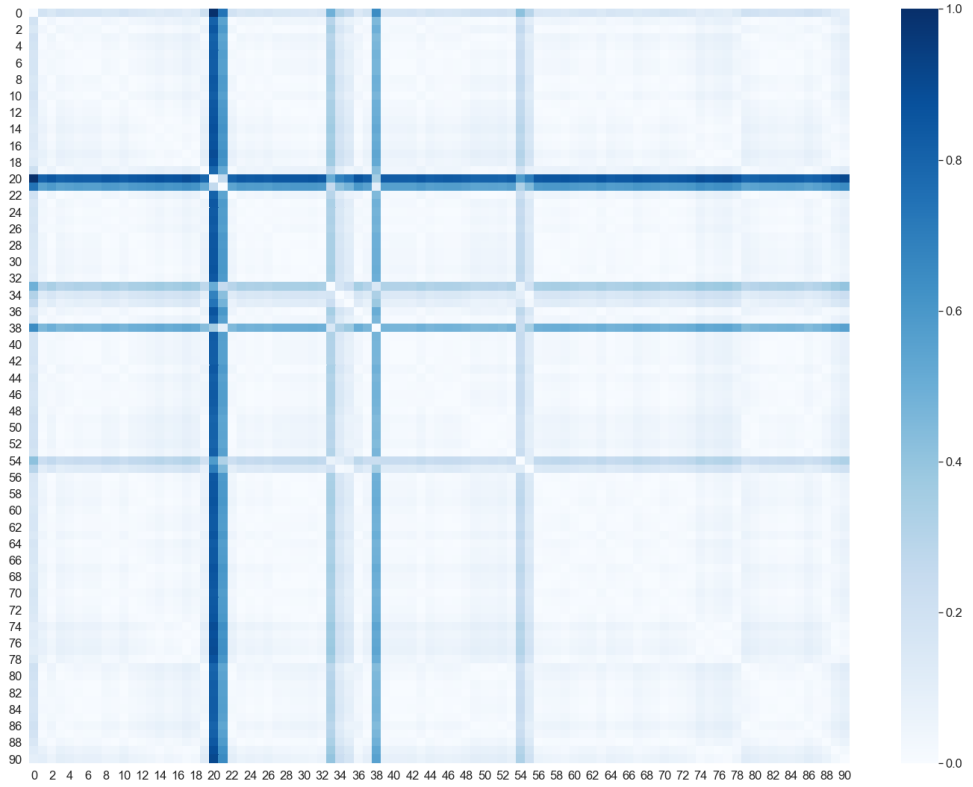


Figura 15: Exemplificação da matriz de distância gerada pela análise da variável P3.

Para identificar os *outliers*, a partir da matriz de distância D obtida, são calculadas as distâncias médias m_i para cada um dos registros i, conforme mostrado na Equação 13.

$$m_i = \frac{1}{N}\sum_{j=1}^{N} d_{ij} \qquad \text{Equação 13}$$

Em que d_{ij} é a distância entre os registros i e j e N é o total de registros existentes na matriz D.

As distâncias médias m_i são então colocadas em ordem crescente, e os índices i correspondentes aos maiores valores de m_i são os registros considerados como dados anômalos pelo método das distâncias.

O método é exemplificado pela Figura 16. Na figura, após o cálculo das distâncias médias, estas estão em ordem crescente. Para o exemplo, podem ser considerados dados anômalos, os registros com distância média superior a 0,2, que é o

limite de corte para este caso. Os registros assinalados como anômalos são os pontos 20, 21, 33, 38 e 54, que são os pontos com cores mais escuras na Figura 15.

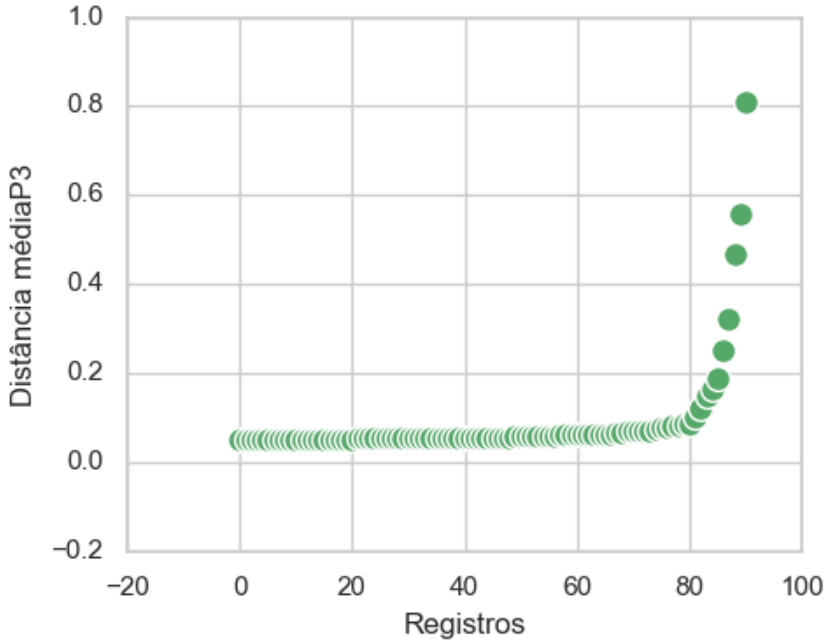


Figura 16: Exemplificação das distâncias m_i distribuídas em ordem crescente para variável P3.

O procedimento descrito utilizando as distâncias médias dos registros foi aplicado de forma univariada, analisando cada uma das variáveis individualmente, e em uma análise multivariada, considerando simultaneamente todas as variáveis. Além disso, é importante mencionar que na análise multivariada, antes do cálculo da matriz de distância dos registros, os dados de cada uma das variáveis foram padronizados, utilizando a normalização Min-Max, mostrada na Equação 14.

$$y_i^j = \frac{x_i^j - min_j}{max_j - min_j} \qquad \text{Equação 14}$$

Em que min_j e max_j representam respectivamente o valor mínimo e máximo da variável j para o conjunto de registros i analisados (AGGARWAL, 2015).

A necessidade de padronização se dá em função das variáveis estarem em diferentes escalas, e em função disso, não podem ser comparadas entre si. Dessa forma, sem a padronização, as variáveis com maior magnitude irão influenciar mais os resultados nos cálculos das distâncias.

O último método empregado na análise de pontos anômalos é o algoritmo LOF (*Local Outlier Factor*) (BREUNIG et al., 2000). Este método é baseado no conceito de densidade local, em que a densidade de um registro é comparada com as densidades de seus k vizinhos mais próximos. Assim, pode-se identificar regiões de densidade similar e os pontos com densidade substancialmente inferior aos seus vizinhos, que são os pontos considerados como *outliers*. O fator LOF é uma pontuação que indica se um determinado registro é um dado anômalo ou não. De forma geral, valores de LOF próximos de 1 são considerados normais, e medidas de LOF muito superiores a 1 são consideradas *outliers*.

Primeiramente, para implementar o algoritmo, é necessário padronizar os dados das variáveis, feito aplicando o método Min-Max mostrado na Equação 14, e após isto calcular a matriz de distância entre os registros, que também para este método se utilizou a distância euclidiana, mostrada na Equação 13. Outro dado de entrada importante é o número de vizinhos k que serão considerados na análise.

Com a matriz de distância e o número de vizinhos definidos, o primeiro passo é calcular, para cada um dos registros, sua $\mathrm{k-distância}_o$, ou seja, a distância entre o ponto em análise e o seu k-vizinho mais distante. Assim, por exemplo, considerando o registro O e que a análise é feita considerando 8 vizinhos, o valor de k-distância para o ponto O será a distância entre o ponto O e seu oitavo vizinho mais distante. Após isso, é calculada, para cada um dos registros, a distância de alcançabilidade (ou *reachability distance*), dada pela Equação 15.

$$\mathrm{reach\_dist}_k(p,o) = max\{\mathrm{k-distância}_o, d_{p,o}\} \qquad \text{Equação 15}$$

Considerando que p é o ponto em análise, a distância entre o ponto p e o ponto o é o maior valor entre a medida k-distância do ponto o e a distância entre p e o, $d_{p,o}$. A Equação 15 funciona como fator de suavização para o cálculo da distância entre os registros. Se o ponto p for um dos k vizinhos do ponto o, a distância entre os dois pontos será dada pelo valor de $\mathrm{k-distância}_o$, caso contrário, será a medida normal entre a distância entre os dois pontos.

O próximo passo é o cálculo da densidade da vizinhança de um ponto p, a medida lrd_p, dada pela Equação 16.

$$\mathrm{lrd}_k(p) = \frac{1}{\frac{\sum_{o \in N_k(p)} \mathrm{reach\_dist}_k(p,o)}{k}}$$ Equação 16

A medida $\mathrm{lrd}_k(p)$ é calculada para cada registro p, considerando seus k vizinhos mais próximos. Os pontos pertencentes a vizinhança do ponto p, $N_k(p)$, são utilizados para o cálculo da distância de alcançabilidade. O valor $\mathrm{lrd}_k(p)$ representa o inverso da média das distâncias. Assim, distâncias maiores de um ponto aos seus vizinhos correspondem a densidades menores.

Finalmente, o valor $\mathrm{LOF}_k(p)$ (*local outlier factor*) para cada ponto p pode ser calculado

$$LOF_k(p) = \frac{\sum_{o \in N_k(p)} \frac{\mathrm{lrd}_k(o)}{\mathrm{lrd}_k(p)}}{k}$$ Equação 17

A medida $LOF_k(p)$ indica o grau que o ponto p pode ser considerado um *outlier*, já que compara sua densidade com a média das densidades de seus vizinhos. Caso a medida de densidade de um determinado ponto seja muito inferior à de seus vizinhos, o valor de LOF será muito superior a 1, de forma que o ponto se situa em uma região longe de áreas densas, e por isso, é um forte indicativo de ser um *outlier*.

Assim como feito com o método das distâncias, o algoritmo LOF foi aplicado de forma individual para cada variável e considerando simultaneamente todas as variáveis. Na prática, identificar como dados anômalos todos os pontos com valores de LOF maiores que 1 é uma análise muito conservativa. Por isso, após os valores de LOF calculados, um histograma é gerado, conforme exemplificado na Figura 17, para analisar como as medidas estão distribuídas. No exemplo mostrado na figura, percebe-se que a maior parte dos dados apresentam valores de LOF inferiores a 1.5, por isso, a medida de corte fixada para este caso passa a ser 1.5, e os *outliers* dessa variável serão todos os registros com medidas de LOF superiores a 1.5.

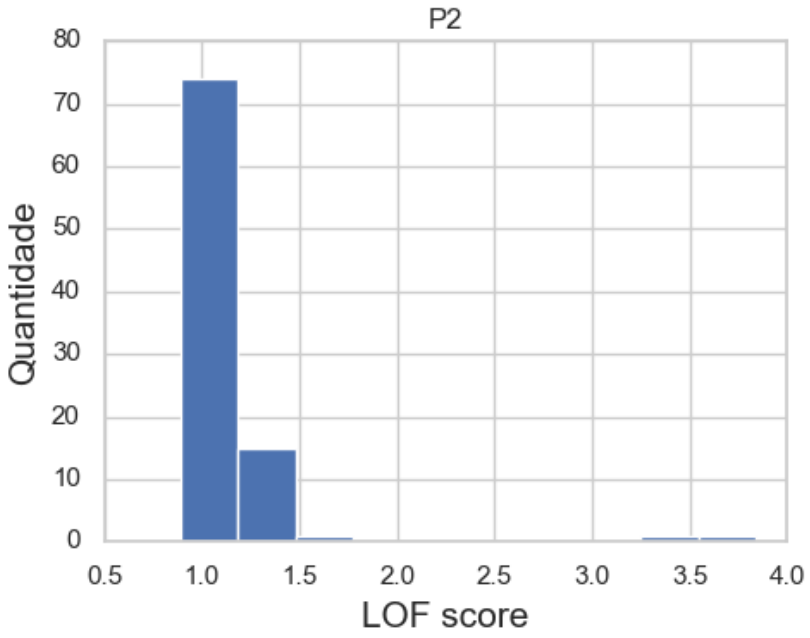


Figura 17: Exemplificação do histograma gerado com as medidas de LOF para análise da variável P2.

Os pontos de corte foram determinados a partir dos histogramas. Para conjunto de dados que estão distribuídos em forma de grupos, o método LOF consegue ter um desempenho superior aos outros dois métodos expostos neste trabalho.

Após a identificação de dados anômalos pelos três métodos, os gráficos correspondentes a cada procedimento são mostrados simultaneamente, conforme exemplificado na Figura 18. Para cada variável analisada, os registros são dispostos ao longo do tempo, e pontos identificados como *outliers* são assinalados com a cor preta, juntamente com o índice correspondente do registro na tabela de dados. Dessa forma, a análise de dados anômalos passa a ser gráfica, avaliando qual método melhor se ajusta para a variável sob análise, ou quais dos pontos registrados sãos potenciais *outliers*. Esta análise deve ser criteriosa, levando também em consideração conhecimentos físicos e operacionais. Dessa forma, os métodos atuam como ferramentas de suporte, para colaborar na determinação dos registros duvidosos.

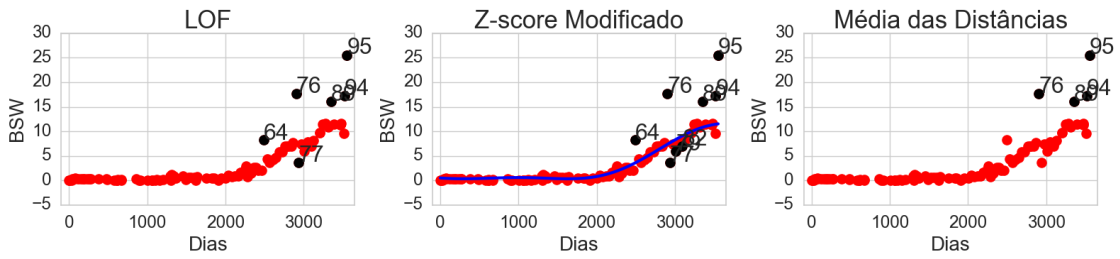


Figura 18: Exemplificação do resultado gerado pela aplicação dos métodos de identificação de *outliers*.

Entretanto, os gráficos gerados, como o do exemplo da Figura 18, só são aplicados para análise individual de cada variável. Dessa forma, para complementar os estudos, um gráfico com um resumo geral das análises é gerado para cada um dos métodos aplicados. A Figura 19 mostra um exemplo desse gráfico gerado para um determinado poço, utilizando o método LOF. No gráfico, as linhas indicam os registros analisados. Cada registro é um teste de produção. As colunas são as variáveis analisadas. Para os métodos LOF e de distância, a última coluna é a análise múltipla das variáveis. Os *outliers* são indicados pelas cores verde e vermelha. Registros em que o teste era inválido e *outliers* foram apontados, são indicados pela cor vermelha, e testes válidos em que o método apontou como sendo *outlier*, são indicados pela cor verde. No exemplo mostrado, todos os pontos apontados como *outliers* pela análise de todas as variáveis (última coluna) são testes inválidos.

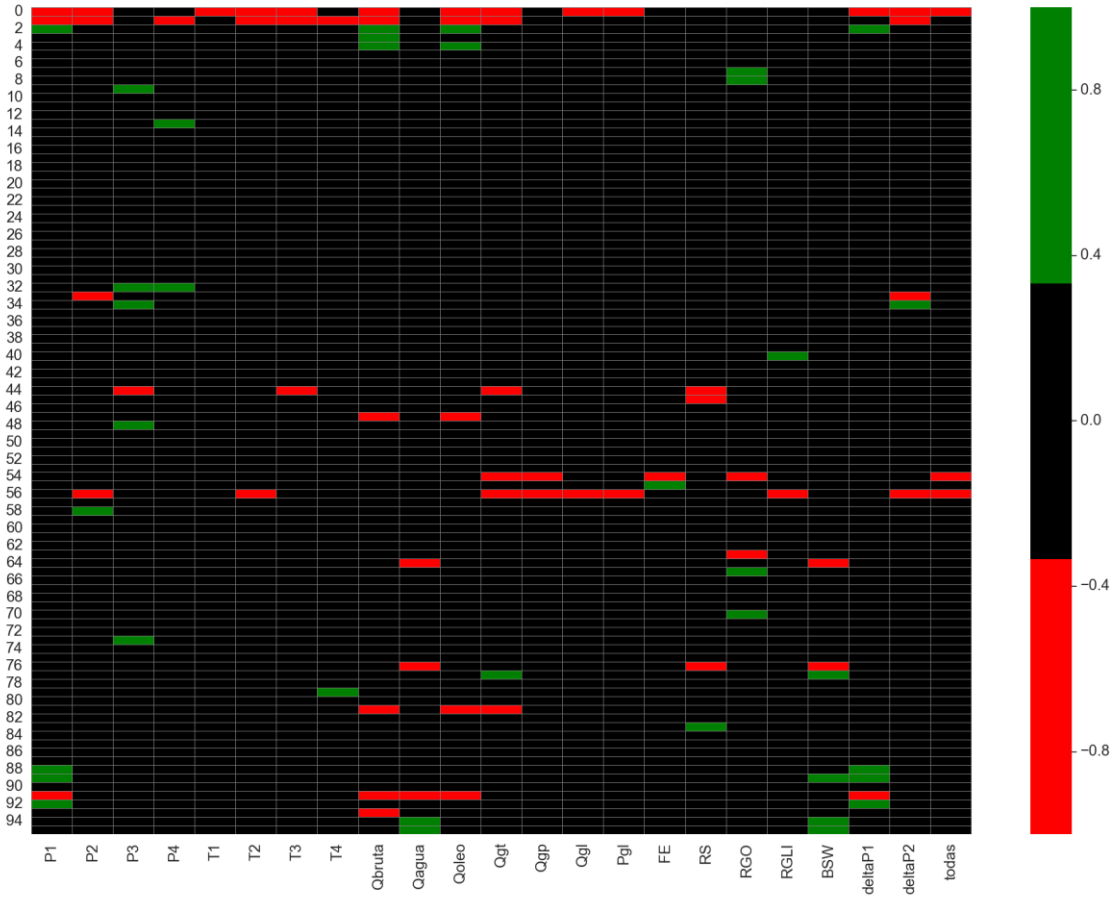


Figura 19: Exemplificação do resumo de resultados obtidos na aplicação do método LOF.

Após uma análise criteriosa dos dados anômalos, estes são corrigidos ou removidos, apenas considerando os testes válidos. Em alguns casos, é possível perceber erros de digitação e estes pontos podem ser corrigidos. Nos casos em que não se sabe a causa do problema, o dado é removido.

Outro problema existente nesta fase de pré-processamento de dados é a presença de dados faltantes, oriundos das planilhas originais e dos dados removidos pela análise de *outliers*. A princípio, considerou-se aplicar algum método para inserção de valores nestes espaços vazios. Entretanto, como muitas variáveis são correlacionadas, inserir dados pontuais poderia trazer problemas para as análises futuras, já que prejudicaria na análise integrada das variáveis para obtenção da variável dependente. Além disso, agregaria mais incerteza ao processo.

Em vista do problema, foram definidas duas formas de lidar com dados ausentes. Nas análises em que uma determinada variável apresentava uma quantidade considerável de dados faltantes, ela era removida da análise. Em uma segunda

abordagem, para as variáveis com ausência de poucos registros, a linha era inteiramente removida, ou seja, o teste era desconsiderado.

5.3 Classificação dos Testes de Produção

Após a conclusão da fase de pré-processamento, a segunda etapa é classificar um teste de produção como válido ou inválido, de acordo com as variáveis de produção disponíveis. O problema é então resolvido como um modelo preditivo de classificação, em que as entradas são as variáveis em análise e a saída é a variável binária Status, mostrada na Tabela 8, que categoriza o teste de produção em válido e não válido. O objetivo é detectar um evento raro, no caso o teste de produção inválido, de forma supervisionada.

Conforme visto, os testes de produção de petróleo são classificados como válidos e não válidos de acordo com a rotulação dos especialistas. Afim de se evitar possíveis confusões em relação a nomenclatura de testes de produção de petróleo, com a nomenclatura dos procedimentos metodológicos dos modelos de classificação na área de aprendizado de máquinas, irá se adotar a seguinte terminologia para testes de produção de petróleo: testes válidos fazem parte da categoria *Sim*, e testes inválidos da categoria *Não*. Dessa forma, os testes de produção serão, a partir de agora, considerados como classe Sim e classe Não.

A metodologia proposta para o estudo da etapa de classificação é mostrada na Figura 20. A partir dos dados tratados, estes são divididos em dois grupos, a série de treinamento e a série de validação. Todos os ajustes necessários são realizados na série de treinamento. Os dados deste conjunto passam então por uma etapa de ajuste inicial, seleção de variáveis e calibração dos modelos. A série de validação é somente utilizada no final do procedimento, para avaliar o resultado obtido. O procedimento será explicado com mais detalhes a seguir.

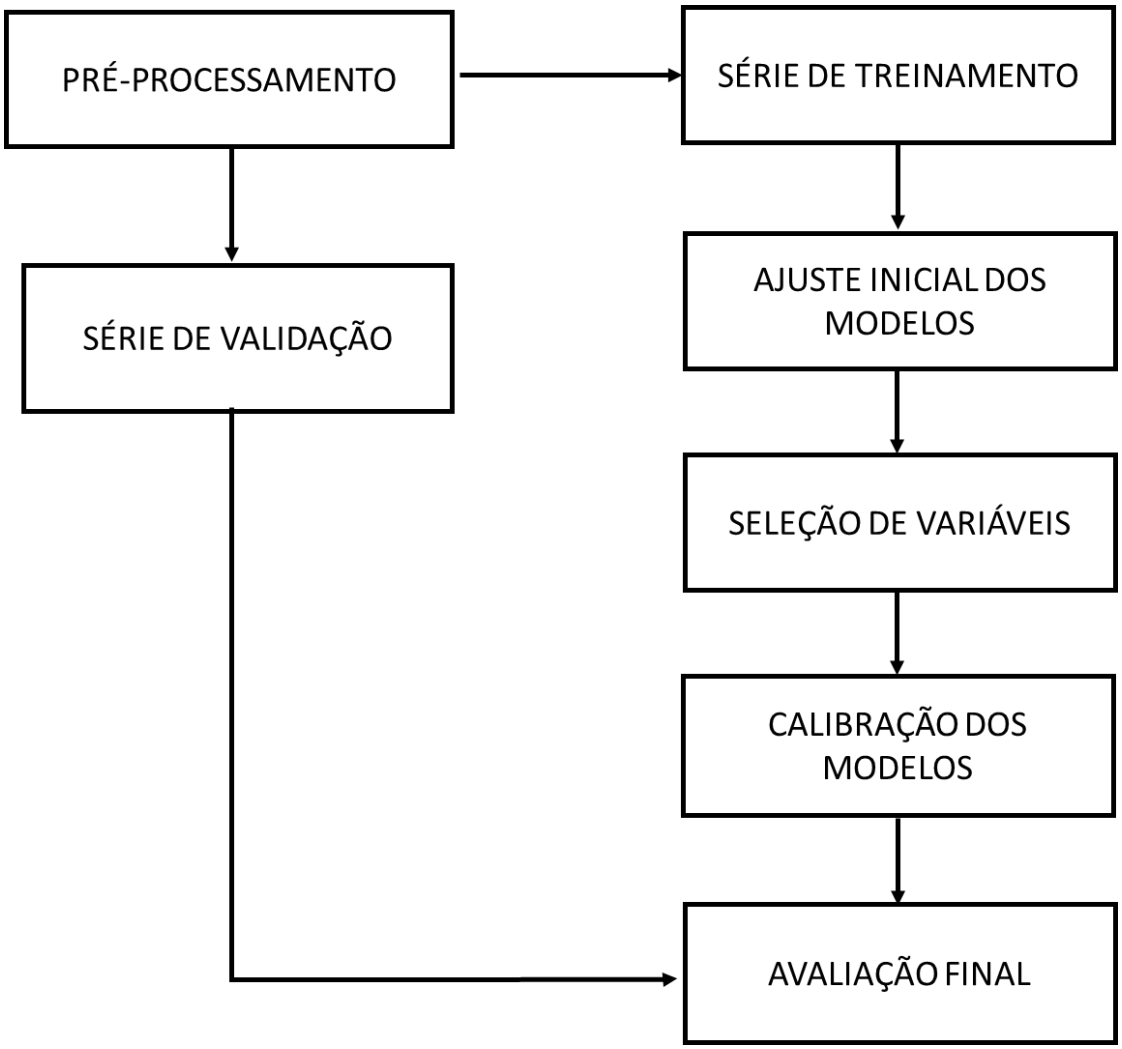


Figura 20: Esquematização da metodologia de classificação de testes de produção.

A parte inicial do estudo consiste em dividir os dados em série de treinamento e série de validação. Para formar o conjunto de validação, são separados os três últimos registros com a categoria tipo *Sim* e os três últimos com a categoria tipo *Não,* do conjunto total de dados disponível. Esta divisão foi feita pensando no objetivo do trabalho de previsão de testes de produção futuros, de acordo com o histórico de dados disponível. Dessa forma, o conjunto de treinamento considerado é o histórico de testes de produção mais antigos e o conjunto de validação são os dados mais recentes.

Todos os ajustes e melhorias nos modelos de classificação são feitos com o conjunto de treinamento. Além disso, procura-se com este conjunto, selecionar quais modelos de classificação são mais adequados para o problema. Para isto, é necessário separar dentro desse conjunto de treinamento, uma parte dos dados para obter as métricas de avaliação que irão analisar o ajuste dos modelos. Entretanto, considerar apenas uma parte dos dados como o único conjunto para testar o ajuste dos modelos

pode levar a problemas de sobre ajuste (*overfitting*). Nestes casos, os modelos treinados se ajustam muito bem a uma pequena amostra, mas desconsideram os ruídos que geralmente estão presentes em dados reais. Dessa forma, estes modelos perdem sua capacidade de generalização, e não conseguem prever eficientemente novos resultados fora deste conjunto de treinamento. Para tentar amenizar este problema, este trabalho utilizou como procedimento para fazer suas avaliações, a validação cruzada de K ciclos.

Na validação cruzada de k ciclos, os dados iniciais são divididos aleatoriamente em k subconjuntos mutuamente exclusivos, D_1, D_2,..., D_k, cada um com tamanho aproximadamente igual. O treinamento e o teste são realizados k vezes. A cada iteração i, a partição D_i é reservada para o conjunto de testes e as partições restantes são utilizadas coletivamente para treinar o modelo (HAN et al., 2012).

Nesta dissertação, foram considerados 10 ciclos para validação cruzada e a métrica AUC (*Area Under the ROC Curve*), mostrada no item 4.2.1.7, foi utilizada para avaliar o desempenho do modelo em cada ciclo. O conjunto de dados analisado é desbalanceado, ou seja, a quantidade de testes da classe *Não* é bem inferior a quantidade de testes da classe *Sim*. Conforme visto, a acurácia não é a medida mais apropriada para esses casos, de forma que a métrica utilizada para avaliação foi a medida AUC, que é uma medida mais robusta para análise de séries desbalanceadas. Dessa forma, para cada rodada, o AUC_i era calculado, e o resultado final considerado era a média das 10 rodadas. A Figura 21 mostra a esquematização do procedimento adotado.

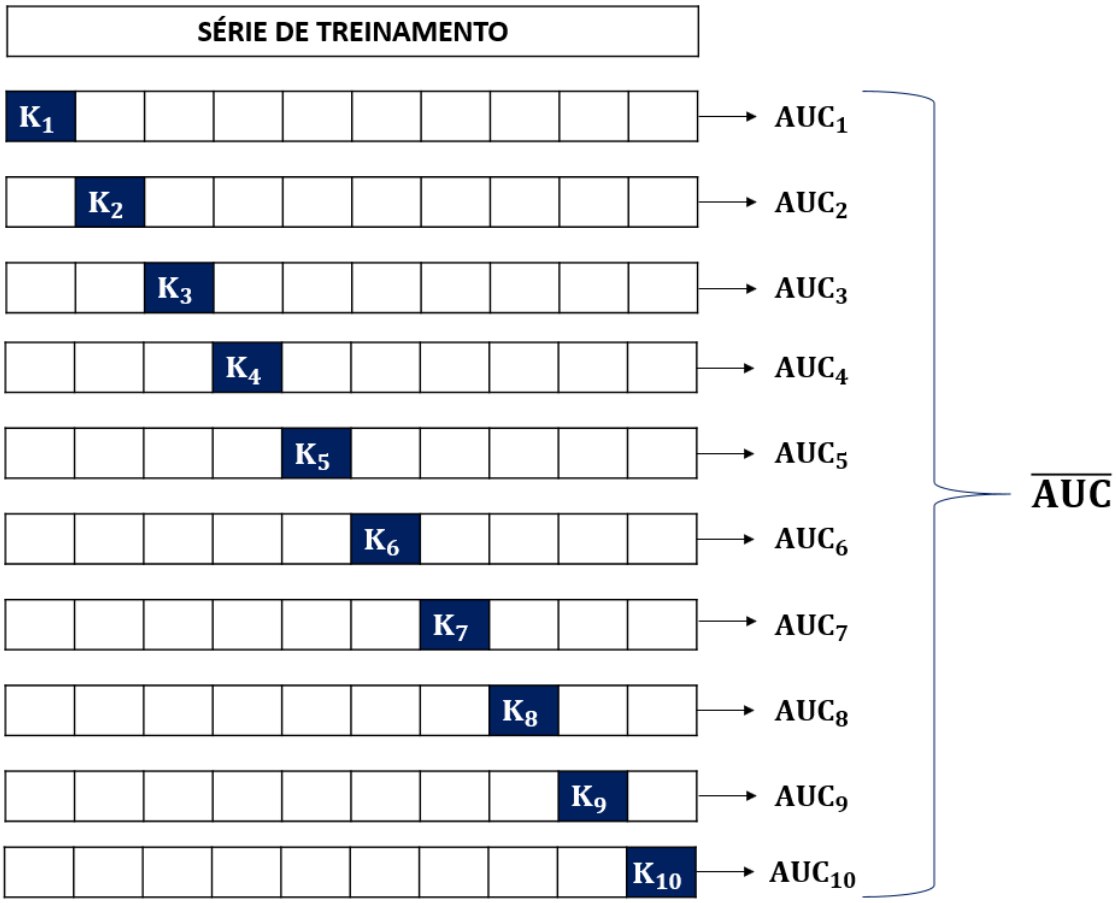


Figura 21: Esquematização da validação cruzada na série de treinamento.

Assim, cada ciclo K corresponde a um conjunto de registros (ou testes de produção de petróleo) escolhidos aleatoriamente. A validação cruzada foi feita de forma estratificada, ou seja, cada ciclo continha a mesma proporção da classe *Sim* e *Não* do conjunto original. Antes de se iniciar o estudo de classificação, os registros de cada ciclo K eram selecionados e armazenados, para que as mesmas amostras de cada ciclo fossem usadas nas etapas subsequentes.

Selecionadas as amostras de cada conjunto k da série de treinamento, é feito o ajuste inicial dos modelos. Foram utilizados os algoritmos dos modelos de classificação oferecidos pelo pacote *scikit learning* (PEDREGOSA et al., 2011), disponível no *Python*. Os modelos de classificação utilizados são mostrados na Tabela 9, e a fundamentação teórica dos métodos pode ser vista no item 4.2.1.

Tabela 9: Modelos de classificação considerados.

Sigla	Modelo
RL	Regressão Logística
NB	Classificador Naive Bayes
KNN	K-Vizinhos mais próximos
DT	Árvore de decisão
RF	Floresta Aleatória
SVM	Máquina de Vetores de Suporte

Cada um dos modelos mostrados na Tabela 9 foi inicialmente ajustado de acordo com seus respectivos parâmetros. A escolha dos parâmetros foi feita observando os resultados das matrizes de confusão e os valores médios de AUC obtidos pela validação cruzada. Inicialmente, a maior parte dos modelos não conseguia prever a classe *Não*, em função do desbalanceamento do conjunto de dados. Por isso, nesta etapa, procurou-se ajustar o peso das classes, colocando um peso maior para a classe com menor número de registros. Também neste ajuste inicial, foi definido os valores de custo para os modelos RL e SVM, a quantidade de vizinhos no método KNN, os critérios de divisão nos nós nos métodos de DT e RF, além da quantidade de árvores utilizadas no RF. Ainda em relação ao modelo não linear SVM, foi utilizado a função núcleo do tipo radial. O processo foi feito para cada poço analisado.

Após definidos os parâmetros de todos os modelos no ajuste inicial, a validação cruzada foi refeita usando as amostras de cada ciclo K, e deste processo, se tem como resultado um valor de AUC médio. Esta etapa de ajuste inicial é nomeada como Etapa 1, e desta se tem os modelos ajustados e o valor médio, chamado $\overline{AUC_1}$, para cada um dos modelos de classificação. Os valores de $\overline{AUC_1}$ são usados para a próxima etapa de seleção de variáveis.

A segunda etapa, correspondente a seleção de variáveis, é nomeada Etapa 2. Ao se analisar um conjunto de dados, muitas das variáveis podem não ser úteis para o ajuste de um determinado modelo de classificação. Além disso, a presença de variáveis redundantes e/ou irrelevantes pode prejudicar o desempenho do modelo. Dessa forma, foi aplicada a Etapa 2 para melhorar a performance dos algoritmos de aprendizado, além de simplificar esses modelos. Existem diferentes métodos de seleção de variáveis na literatura. O método escolhido foi baseado no método *Wrapper* que utiliza o próprio modelo de aprendizado para avaliar a qualidade da seleção dos subconjuntos. Assim, o

objetivo do método é encontrar o melhor subconjunto de variáveis que se adeque ao modelo de classificação. Os modelos de aprendizado de máquinas podem ser melhorados, por exemplo, acrescentando variáveis promissoras em um conjunto vazio, ou retirando as variáveis do conjunto original de dados (JOHN; KOHAVI; PFLEGER, 1994). A Figura 22 mostra uma esquematização do procedimento aplicado.

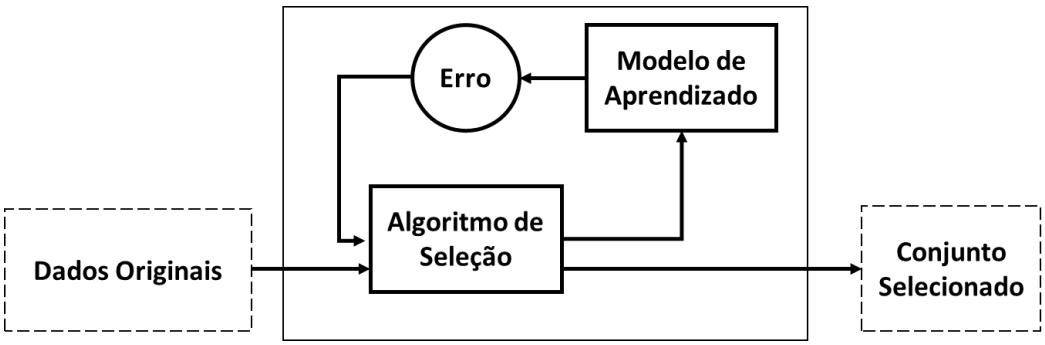


Figura 22: Esquematização do método de seleção de variáveis aplicado.

No procedimento elaborado, conforme mostrado na Figura 22, se utilizou $\overline{\mathrm{AUC}_1}$ como parâmetro inicial de medida de avaliação e, inicialmente são consideradas todas as variáveis. Assim, no módulo *Algoritmo de Seleção*, a variável da posição j=0 é removida, e o subconjunto restante vai para o módulo *Modelo de Aprendizado*, em que o modelo de classificação é rodado usando os K-ciclos da validação cruzada. Deste procedimento, é obtido como resultado a medida média de $\overline{\mathrm{AUC}_j}$ correspondente a rodada j. Esta medida de erro volta para o algoritmo de seleção, em que é feito um teste comparativo entre o valor de $\overline{\mathrm{AUC}_j}$ corrente e o novo valor obtido. Se o valor de $\overline{\mathrm{AUC}_j}$ for maior que o valor corrente, significa que retirar a variável melhorou a qualidade do modelo, e a variável da posição j é removida definitivamente, e o novo valor de $\overline{\mathrm{AUC}}$ corrente é atualizado. Caso contrário, a variável volta para o conjunto original. Após isso, a variável da posição j=1 é removida e o subconjunto restante novamente é conduzido para o módulo *Modelo de Aprendizado* e todo processo é refeito, até que todas as variáveis sejam testadas, e o índice j atinja o valor de número de variáveis total.

O procedimento descrito é aplicado individualmente para cada modelo. Cada modelo irá ter no final deste processo um subconjunto de variáveis que a princípio melhoram sua qualidade, e uma medida $\overline{\mathrm{AUC}_2}$, que é a medida de erro média característica desta etapa 2, de seleção de variáveis. Apesar deste procedimento melhorar a qualidade dos resultados para certos casos e para determinados modelos

avaliados, entende-se que este processo contém limitações. Primeiramente, por utilizar apenas a medida média do parâmetro AUC, e por não otimizar o processo de busca de soluções, fazendo diferentes seleções de variáveis e analisando suas medidas de erro. Além disso, para modelos como DT e RF, que não são robustos, este procedimento pode não ter efeito.

Após fazer a análise 2 para cada um dos modelos de classificação, o processo de validação cruzada com os K ciclos foi refeito para cada um desses modelos, para verificar situações de falta de robustez dos procedimentos. Isto ocorre especificamente nos casos de RF, em que a seleção de variáveis e registros nos nós se dá de forma aleatória. Então, para alguns casos, o valor médio de $\overline{\mathrm{AUC}_2}$ pode ser ligeiramente inferior que $\overline{\mathrm{AUC}_1}$.

A etapa 3 refere-se à calibração dos modelos, utilizando as variáveis de cada modelo obtidas pela etapa 2. Este procedimento consiste em testar diferentes combinações de parâmetros específico de cada modelo de aprendizagem, para otimizar seu desempenho. Esta procura dos melhores parâmetros está associada ao modelo de classificação considerado, ao espaço de parâmetros desse estimador, o método de busca ou candidatos de amostras e a métrica de erro que irá analisar a qualidade dos resultados. A análise foi feita utilizando a função *GridSearchCV* do pacote *scikit learning*. O método considera exaustivamente todas as combinações de parâmetros disponibilizada, utilizando o esquema de validação cruzada e uma métrica para avaliar o desempenho da combinação de parâmetros em cada validação. Desta função são obtidos como resultados o melhor conjunto de parâmetros e a medida de erro obtida. Utilizou-se como medida de erro o valor de AUC. Os parâmetros considerados para análise são mostrados na Tabela 10.

Tabela 10: Parâmetros otimizados em calibração dos modelos.

Método:	Parâmetros:	Valores
RL	C	0.0001, 0.001, 0.01, 1, 100
	Penalidade	'l1', 'l2'
KNN	Número de Vizinhos	De 1 até 31
DT	Profundidade Máx	De 1 a 20, com passo 2
	Critério	Gini ou Entropia
RF	Número de estimadores	10, 20, 30, 50, 100
	Critério	Gini ou Entropia
SVM	C	10^{-3} até 10^{7}
	Gama	10^{-5} até 10^{3}

Após a finalização desta etapa, a validação cruzada é novamente refeita com os K-ciclos, considerando nesta fase o subconjunto de variáveis e os parâmetros otimizados obtidos pelo *GridSearchCV*. Deste processo, obtém-se a medida média de AUC, $\overline{\mathrm{AUC}_3}$. A série de treinamento é então ajustada com os parâmetros otimizados e o subconjunto de variáveis próprios de cada modelo, e o melhor ou melhores modelos são utilizados para prever as classes da série de validação, de acordo com as entradas do conjunto de validação.

5.4 Previsão das Variáveis do Processo

Após a finalização da etapa de classificação, a etapa posterior deste estudo é a previsão das variáveis do processo. O objetivo desta fase é encontrar um valor previsto e um intervalo de predição robustos para cada variável. Dessa forma, supondo que estes intervalos são confiáveis, caso uma variável obtida durante o teste de produção esteja fora deste intervalo, pode ser um indicativo que esse teste de produção apresenta algum problema ou que a variável modificou seu comportamento ao longo do processo de produção. O problema de previsão de variáveis é resolvido como um modelo de regressão que utiliza o histórico de produção para estimar o valor esperado da variável e seu intervalo na data que o teste será realizado.

Muitas das etapas de previsão das variáveis do processo são similares com as de classificação mostrado no esquema da Figura 20. As diferenças principais desta etapa estão na consideração de apenas testes de produção da categoria *Sim* para análise, e na

obtenção dos resultados de previsão de forma estocástica. As etapas da previsão das variáveis serão vistas com detalhes abaixo.

Conforme mencionado, nesta etapa se trabalha apenas com testes de produção válidos (ou seja, os testes da categoria *Sim*). A justificativa é que se um teste de produção foi considerado válido pelo especialista, acredita-se que os valores obtidos nas variáveis principais sejam coerentes, dentro de um intervalo esperado. E não se tem essa confiabilidade para os testes de produção categorizados como *Não*. Assim, como um passo inicial da fase de previsão de variáveis está a filtragem dos dados da categoria *Sim*. Para estes dados, conforme feito no estudo de classificação, são separados os conjuntos de treinamento e de validação. A série de validação neste caso é formada pelos 4 últimos registros, ou seja, os 4 últimos testes de produção.

Outra característica importante são as variáveis dependentes consideradas no estudo. Como o principal objetivo dos testes de produção é definir as vazões produzidas por cada poço, as variáveis consideradas como resposta para o modelo de regressão são vazão de óleo, água e gás total. Assim, utilizando os outros parâmetros e variáveis característicos do processo de testes de produção, o objetivo é prever as vazões e construir um intervalo de predição apropriado para cada uma delas.

Também para esta etapa é feita a validação cruzada e os registros são divididos em 10 grupos, correspondente a cada ciclo K da série de treinamento. Selecionadas as amostras, é realizado o ajuste dos modelos. Os modelos de regressão utilizados neste trabalho são mostrados na Tabela 11, e a fundamentação teórica para os modelos pode ser vista no item 4.2.2. Vale ressaltar que, assim como na etapa de classificação, o modelo SVR foi ajustado considerando um núcleo do tipo radial.

Tabela 11: Modelos de regressão considerados.

Sigla	Modelo
MLR	Regressão Linear Múltipla
SVR	Regressão por Vetores de Suporte
RT	Árvore de Regressão
RFR	Floresta Aleatória para Regressão

As etapas de ajuste inicial dos modelos e seleção de variáveis são semelhantes ao procedimento metodológico do modelo de classificação. Os modelos foram ajustados usando os parâmetros padrões do pacote *scikit learning* e a métrica RMSE (mostrada no

item 4.2.2.4) foi utilizada para a análise da validação cruzada nos 10 ciclos. Assim, a média do valor de RMSE obtidos para cada modelo era usada para selecionar quais variáveis considerar em cada modelo.

A etapa de calibração de modelos também foi realizada, mas apenas no modelo SVR. Era muito computacionalmente custoso realizar esta etapa para RFR, e este método, conforme será visto, não tem um desempenho satisfatório para o caso estudado. Para realizar a otimização dos parâmetros do método SVR foi novamente aplicando a função *GridSearchCV* do pacote *scikit learning*. Os intervalos dos parâmetros considerado para esta análise podem ser vistos na Tabela 12.

Tabela 12: Parâmetros otimizados em calibração dos modelos.

Método:	**Parâmetros:**	**Valores**
SVR	C	10^{-1} até 10^{3}
	Gama	10^{-7} e 10^{-4}
	Epsilon	0.01,0.1,0.2,0.5,0.3,0.4, 0.5,0.6,0.7,0.8,0.9,1,100

Sumarizando as etapas, na série de treinamento são separadas amostras em 10 ciclos para validação cruzada, estes ciclos são usados para o ajuste inicial, seleção das variáveis dos modelos de regressão, e finalmente o modelo SVM é otimizado aplicando o método de busca *GridSearchCV*. Os modelos calibrados são então aplicados nas etapas posteriores utilizando a série de validação do modelo de regressão.

Com os modelos calibrados, o objetivo com a regressão é determinar para um novo ponto, o valor da variável dependente a partir dos registros de entrada recebidos. Entretanto, este valor é um valor ajustado, que considera a tendência dos dados, mas não considera os ruídos que são comuns na realidade. Por isso, além de um valor ajustado previsto, é importante levar em consideração o comportamento estocástico do fenômeno natural. Além disso, no problema estudado, é necessário encontrar um intervalo confiável para as previsões, para que, a partir das variáveis obtidas no teste de produção, determinar se estas estão de acordo com seu comportamento esperado. Em uma análise determinística, dificilmente o valor previsto seria o valor real. Para considerar a incerteza do processo de previsão de um novo ponto, ou seja, o erro de previsão do modelo somado ao ruído dos dados de entrada, foi aplicado o método *Bootstrap* (EFRON, 1979) para geração dos intervalos de predição do modelo.

O método de simulação *Bootstrap* é uma técnica de amostragem não paramétrica. A característica principal do método é que amostragem é feita com repetição, ou seja, um novo conjunto de dados é formado a partir de um conjunto original, podendo existir neste novo conjunto termos repetidos. O método *Bootstrap* é aplicado em diferentes áreas. Dentre elas, está a criação dos intervalos de predição, conforme mostrado em STINE (1985). A metodologia elaborada para geração de amostras da variável prevista a partir de um novo dado de entrada X_i, utilizando o processo de *bootstrap* é mostrada na Figura 23.

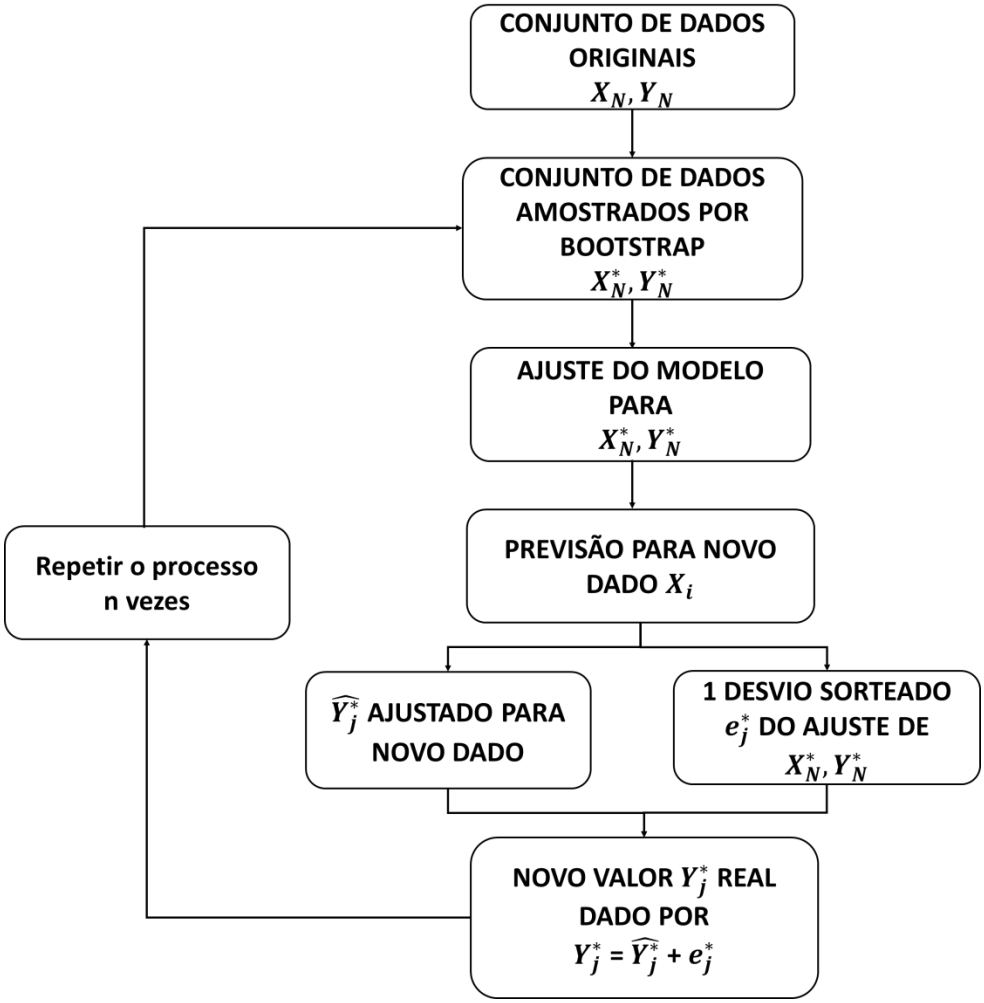


Figura 23: Esquema de geração de amostras da variável dependente pelo método de *bootstrap*.

Na metodologia proposta, partindo do conjunto original da série de treinamento, o processo mostrado na Figura 23 é feito para a primeira data da série de validação. Um conjunto de registros são selecionados pelo método *bootstrap*, gerando uma matriz de dados amostrada de mesmo tamanho do conjunto original. Cada um dos modelos de

regressão é ajustado com essas amostras, e deste ajuste são obtidos: o valor da variável dependente previsto para essa primeira data e um desvio do ajuste sorteado aleatoriamente. No caso de escolha do desvio, é também aplicado a técnica de amostragem *bootstrap*, considerando que o desvio sorteado é obtido a partir do conjunto de desvios do modelo ajustado. O valor previsto é somado com o desvio, e se tem um ponto de uma variável dependente para aquela data. O processo foi repetido 10000 vezes, de forma que no final se tem um conjunto de 10000 variáveis dependentes consideradas possíveis soluções para a data prevista. Deste conjunto de dados, são então calculados os percentis P10, P50 e P90. O intervalo de predição é então formado considerando os limites inferiores e superiores, os percentis P10 e P90, respectivamente. Isto garante que 80% dos dados gerados da variável estão neste intervalo. Assim, nos testes feitos com a série de validação, quanto mais próximo o valor real está do P50, maior a confiabilidade de que o modelo de regressão ajustado.

Diferentemente da classificação em que os testes da série de validação foram previstos simultaneamente, na regressão, após a previsão de cada teste da série de validação, estes eram deslocados para o conjunto de treinamento. Assim, inicialmente o conjunto de validação continha 4 datas para serem previstas, e finalizavam com nenhum dado.

6 ESTUDO DE CASO

Neste capítulo, o procedimento metodológico mostrado no capítulo 5 é experimentado e avaliado para resolver o problema proposto de validação de testes de produção em tempo real. Foram utilizados testes de produção de petróleo de um campo representativo brasileiro. Deste campo, são analisados 13 poços de petróleo com um histórico de produção de, em média, 10 anos, todos no mesmo intervalo de tempo considerado. Todos os poços estudados são produzidos em uma mesma plataforma. Nesta unidade de produção, existe apenas uma planta de teste com um separador trifásico. Dessa forma, os testes de produção são feitos, de tempos em tempos, em cada um desses 13 poços, como também nos demais poços ligados a essa plataforma. O método de elevação artificial utilizado na produção dos 13 poços é o *gas lift*.

Em relação à produção, os poços apresentam comportamentos distintos, com intervalos de vazão de óleo, água e gás diferentes. A Tabela 13 mostra os valores de importantes variáveis analisadas, obtidos do teste de produção válido mais recente para cada um dos 13 poços analisados. O objetivo é mostrar a ordem de grandeza das variáveis analisadas, e como as vazões e parâmetros importantes podem variar de poço para poço. Além disso, é possível verificar na Figura 24 a disposição dos poços de acordo com sua produção de água e de óleo. É possível visualizar que, no momento analisado, cada poço representa uma situação específica de produção, sendo W1 o poço com maior produção de óleo, e W8 o poço com a menor produção.

Tabela 13: Resumo das variáveis características dos 13 poços.

Poços	Qágua (m^3/d)	Qóleo (m^3/d)	RGO (m^3/m^3)	RGLI (m^3/m^3)	BSW (%)	Núm. Testes Válidos	Núm. Testes Inválidos
W1	990	767	88	117	56	78	20
W2	168	378	100	378	31	63	16
W3	336	654	111	220	34	68	13
W4	63	475	103	393	12	71	25
W5	1985	433	81	95	82	60	21
W6	1346	688	77	129	66	76	15
W7	384	571	105	225	40	81	14
W8	152	214	93	531	42	78	24
W9	985	424	84	158	70	73	20
W10	1100	449	102	142	71	76	22
W11	2158	451	68	75	83	71	24
W12	2397	572	151	78	81	70	25
W13	916	318	106	158	74	58	21

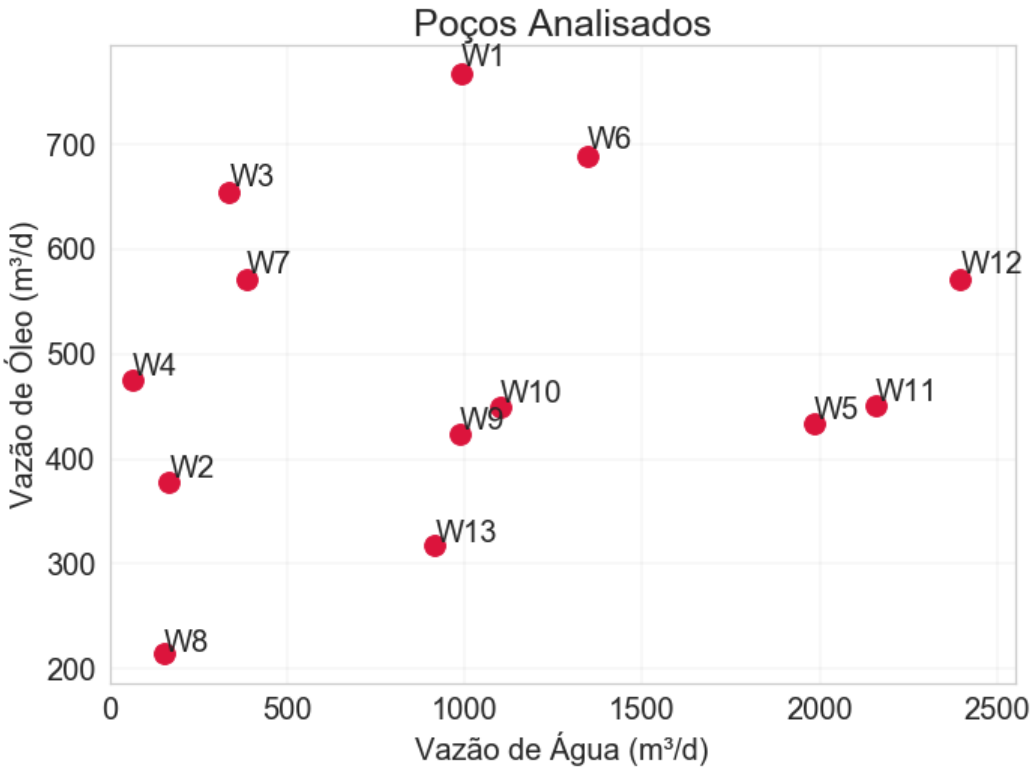


Figura 24: Situação dos 13 poços em relação a vazão de óleo e de água.

A análise dos poços é feita de forma individual. As etapas pré-processamento, classificação e previsão dos dados foram conduzidas para cada um dos poços e os resultados e análises gerais obtidos são expostos neste capítulo.

6.1 Pré-Processamento dos Dados

Nesta etapa, os métodos Z-score modificado, Média das distâncias e LOF foram aplicados para identificação dos dados anômalos. De acordo com os gráficos obtidos dessas análises, foi possível perceber que os principais problemas de pré-processamento ocorrem nas medidas de pressão, principalmente P1 e P2, que são respectivamente as pressões no fundo do poço e na cabeça do poço. Em todos os poços, pelo menos um registro continha ΔP1 (ΔP1 = P1 – P2) e/ou ΔP2 (ΔP2 = P2 – P3) negativo, e os testes de produção com este comportamento não eram necessariamente classificados como inválidos. Este é por exemplo o comportamento do poço W9, conforme pode ser visto na Figura 25, em que os pontos com medidas de ΔP1 negativos são testes de produção considerados válidos. Na Figura 26 é possível verificar as pressões de P1 e P2 para este mesmo poço W9. Percebe-se que o problema dos pontos com medidas negativas de ΔP1 são em função de problemas nas medições de P1, que estão muito abaixo do comportamento esperado.

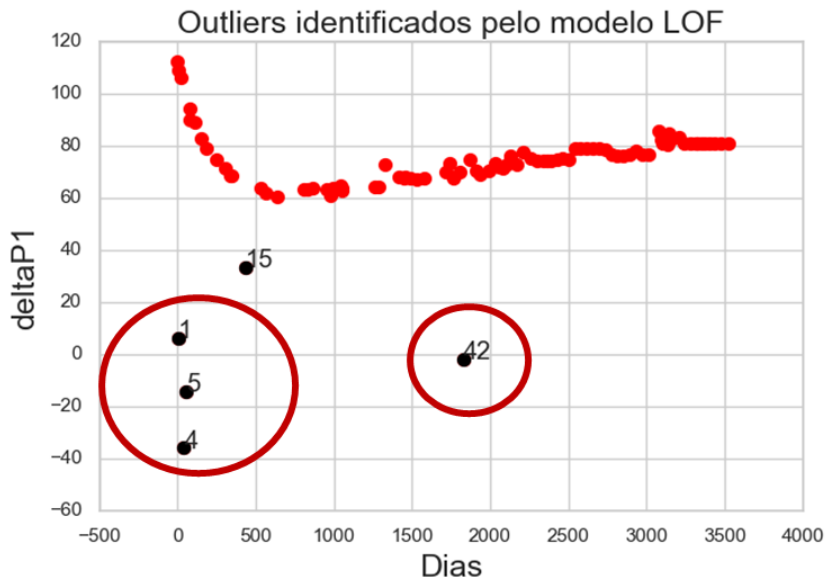


Figura 25: *Outliers* identificados para ΔP1 pelo método LOF para o poço W9.

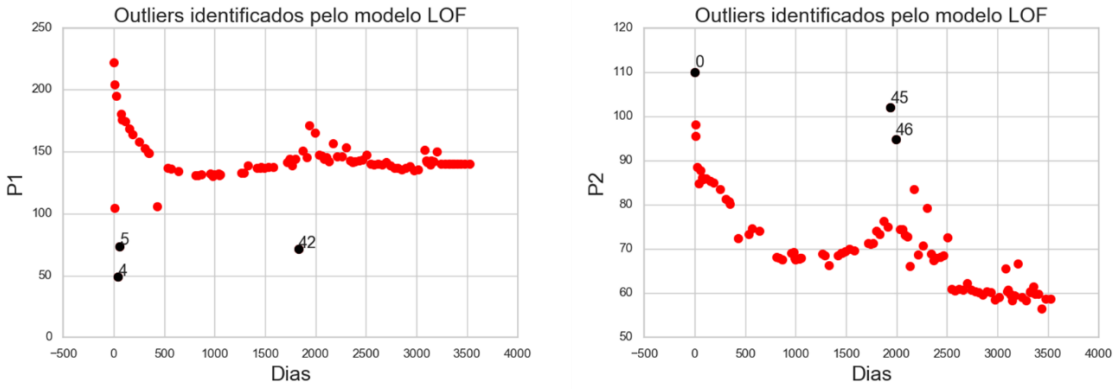


Figura 26: P1 e P2 identificados pelo método LOF para o poço W9.

Outro problema também existente em muitos poços, neste caso principalmente na medida de P1, é a grande quantidade de testes de produção com valores iguais a zero. Este problema é identificado no poço W13, conforme pode ser visto na Figura 27, em que o conjunto dos primeiros pontos identificados como *outliers* (pontos circulados na figura) na medida de ΔP1 são em função de dados inexistentes de P1 no conjunto original.

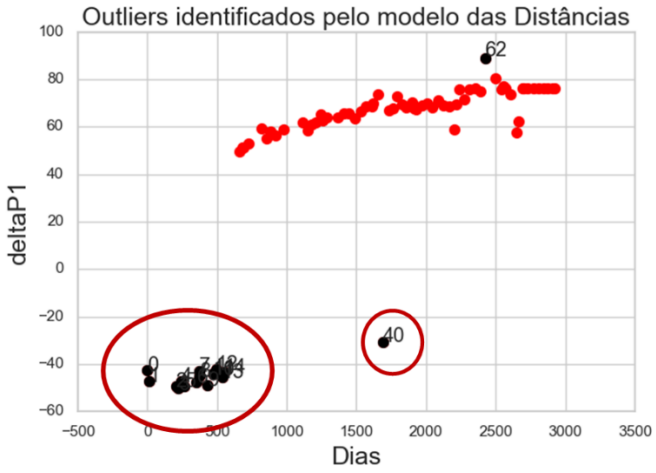


Figura 27: *Outliers* identificados para ΔP1 pelo método LOF para o poço W13.

Para os casos individuais como o da Figura 25 é viável remover os registros, ou linhas da tabela de análise, já que são poucos pontos. Entretanto, os casos como o da Figura 27 são mais complexos, pois remover as linhas significa retirar uma grande quantidade de informação. Para estes casos com muitos registros faltantes é necessário remover a variável. Em função disso, a partir do conjunto de dados disponíveis, as

variáveis de IP, Pe e T4 (descritas nas Tabela 6 e Tabela 8) não foram consideradas nas análises posteriores em função dos longos períodos sem a existência de registros.

Ainda sobre as medidas de pressão, é necessário recomendar sobre o cuidado que operadores devem ter com a utilização das pressões vinda dos testes de produção, em vista da falta de confiabilidade na medida. Mesmo para testes válidos, estas medidas podem conter erros. Neste trabalho, estes dados anômalos foram removidos nesta etapa de pré-processamento, e sugere que o mesmo seja feito na prática normal das atividades. A utilização dos dados brutos sem este cuidado pode ocasionar problemas nos modelos computacionais que utilizam as medidas de pressão dos testes de produção.

Em relação aos métodos de identificação de *outliers*, os três conseguiram identificar dados anômalos eficientemente para diferentes situações de distribuição dos dados ao longo do tempo. De forma geral, o método LOF consegue identificar dados isolados. Entretanto, em situações como mostradas na Figura 27, em que os *outliers* estão agrupados em um *cluster*, dependendo do tamanho do grupo, o método pode ser ineficiente. Neste caso específico, LOF não identificou os *outliers* apontados pelo método das distâncias.

O modelo Z-score modificado tem um excelente desempenho, mas é dependente do modelo de regressão. A Figura 28 mostra um caso em que este método não teve um bom funcionamento. Em função do *cluster* de dados anômalos na parte superior do gráfico, a tendência do modelo de regressão foi deslocada, e dados da parte inferior que não são outliers, foram apontados como sendo. Neste caso específico, Z-score indicou erroneamente dados anômalos que na realidade não são. Entretanto, na análise da Figura 29, na identificação dos dados anômalos para a vazão de óleo no poço W2, Z-score foi o método que apresentou melhor desempenho.

O método da distância é geralmente mais eficiente para modelos que não tem muita variação de comportamento ao longo do tempo, como o caso das pressões e temperaturas. Geralmente para variáveis como as vazões, o método não consegue identificar bem os dados anômalos.

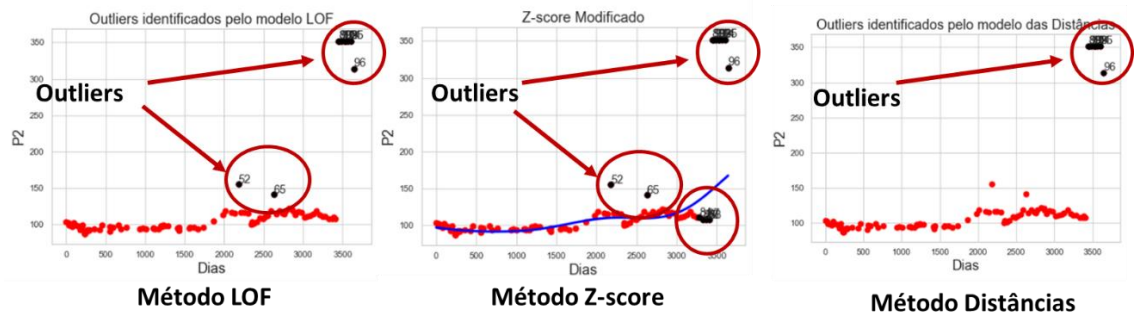


Figura 28: Resultado dos *outliers* identificados para variável P2 do poço W1.

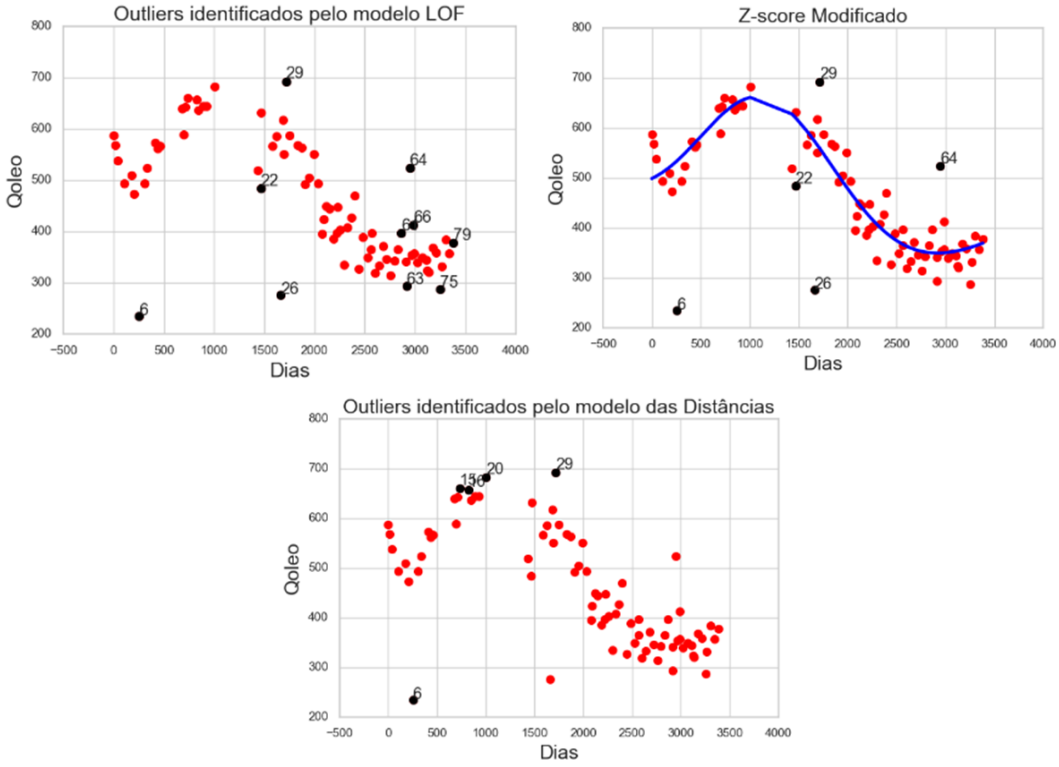


Figura 29: Resultado dos *outliers* identificados para variável Qóleo do poço W2.

A análise conjunta dos três métodos, a partir da análise dos gráficos gerados, permite a identificação dos dados anômalos apropriados a cada variável. Além disso, os métodos LOF e de distância procedem uma análise conjunta com todas as variáveis, de forma que os métodos conseguem identificar registros totalmente irregulares. Para estes casos, os registros eram inteiramente removidos.

6.2 Classificação dos Testes de Produção

De acordo com a metodologia proposta, três etapas principais são seguidas na série de treinamento para a análise da classificação dos testes de produção, mostradas na Figura 30. Os resultados obtidos por cada uma das etapas para o conjunto de treinamento são mostrados neste item. Vale ressaltar que os passos são progressivos, dessa forma, na etapa 3, na fase de calibração dos modelos, a variáveis já foram selecionadas, assim, os modelos são calibrados considerando o subconjunto de variáveis obtidas pela etapa 2.

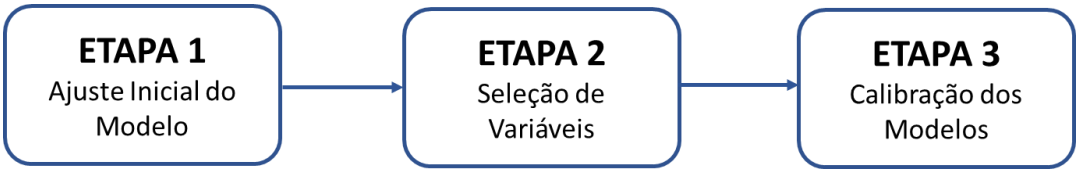


Figura 30: Etapas do processo metodológico do estudo de classificação.

A análise das três etapas mostradas na Figura 30 foi conduzida em todos os treze poços. O poço W1 foi escolhido para mostrar como o estudo foi realizado. A Figura 31 mostra os resultados obtidos do parâmetro AUC e da medida de acurácia para as 3 etapas. Conforme visto na metodologia de classificação, a validação cruzada é feita usando 10 ciclos. Os *boxplots* da Figura 31 foram gerados a partir das medidas obtidas na avaliação de cada ciclo da validação cruzada. Como são os mesmos conjuntos de amostras em cada ciclo, é possível comparar o desempenho das três etapas. Assim, se tem 10 medidas de AUC e de acurácia para cada fase. Apesar da acurácia não ser a medida considerada na avaliação dos resultados, esta também é mostrada para exemplificar o porquê dessa medida não ser adequada para séries desbalanceadas.

A partir da Figura 31, algumas análises dos modelos são feitas. O modelo RL melhora seu desempenho da etapa 1 para etapa 2, a média do AUC é deslocada para direita, mas o modelo é ainda ineficiente e instável, pois tem grande variabilidade nos resultados de AUC obtidos. Na etapa 3, o modelo RL se desloca ligeiramente para direita, próximo do 1, mas continua instável, podendo atingir desde valores bem baixos de AUC até valores razoáveis. Ou seja, a qualidade do modelo depende das amostras analisadas. No modelo KNN, as etapas não trouxeram nenhum impacto. Muito provavelmente o método não conseguiu classificar corretamente o teste tipo *Não*. Um ponto interessante é observar a acurácia do KNN, que é a mesma para todas as etapas, e

está entre 0,7 e 0,8. Esta é a proporção de testes tipo *Sim* nos ciclos, que é a proporção que o método está conseguindo prever corretamente. Por sua vez, o método NB melhorou com a seleção das variáveis. Antes seu AUC estava entre 0,3 e 0,5, e na etapa 2 foi deslocado para a direita, mais próximo de 1. Mas o limite inferior ainda está com um desempenho ruim. A etapa 3 é semelhante a 2, pois o NB não tem parâmetros calibrados nesta fase. O método SVM melhorou seu desempenho da etapa 1 para a etapa 2, mas o método ainda é muito instável, com valores abaixo de 0,5. O desempenho do método SVM na etapa 3 piorou.

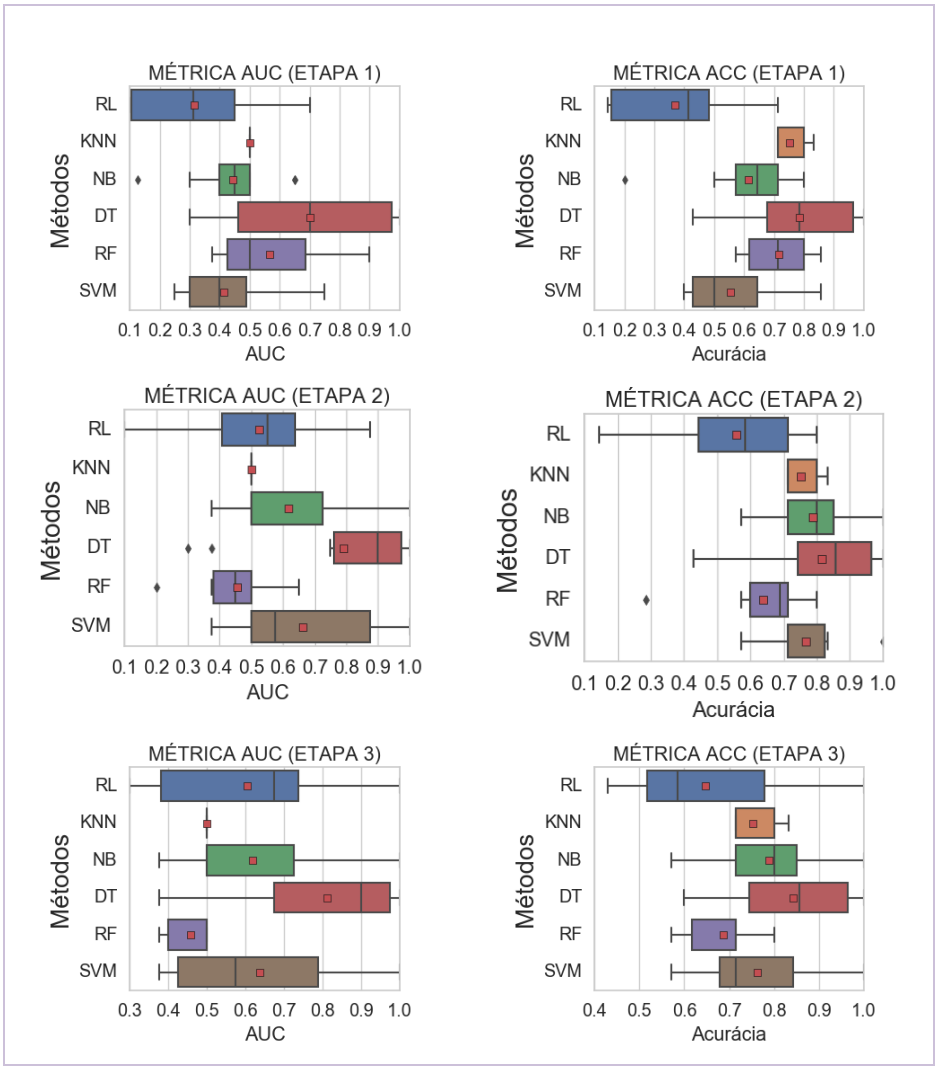


Figura 31: Métricas de AUC e acurácia (ACC) para série de treinamento do poço W1.

Os métodos RF e principalmente o DT foram muito instáveis na análise deste poço e de todos os outros poços analisados. Os dois métodos não são robustos, e as respostas podem variar muito de um caso para outro, de uma semente para outra. No caso do DT, por exemplo, a etapa 2 teve um resultado muito perto de 1, o que poderia ser um excelente modelo. Mas na prática o método pode fornecer qualquer resultado.

Uma outra análise realizada foi avaliar a média e o desvio-padrão do AUC obtidos em cada fase para cada método. O gráfico obtido é mostrado na Figura 32.

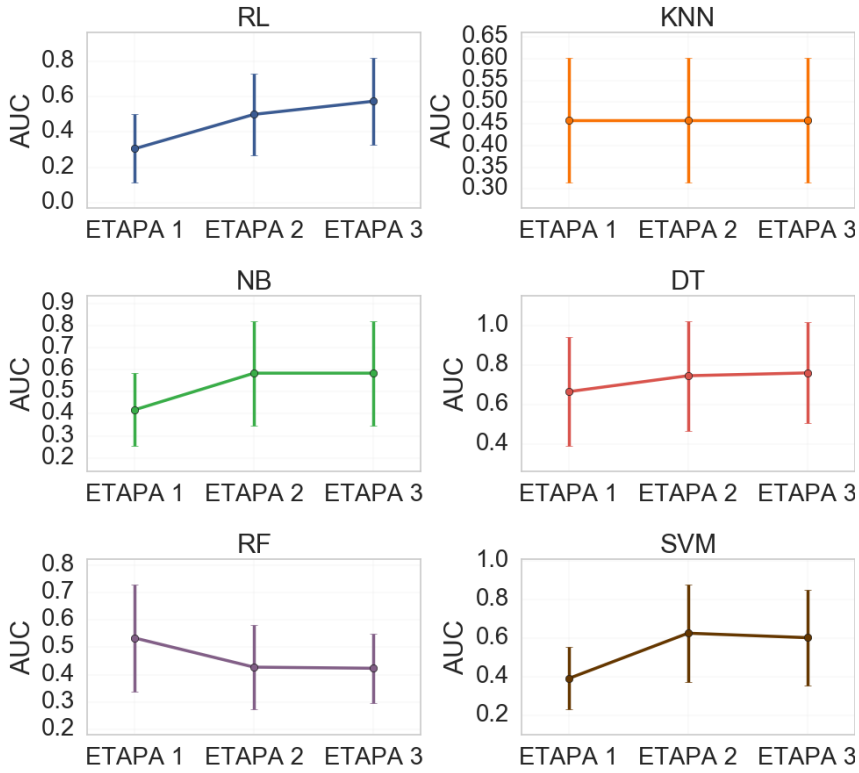


Figura 32: Médias de AUC com desvio-padrão obtidas em cada etapa da série de treinamento para poço W1.

Observando a Figura 32, pode-se perceber que o desempenho da RL melhorou nas etapas, mas seu limite inferior é ainda baixo. A média também é ineficiente, logo o método parece ser ineficiente. O método NB também melhora seu desempenho da etapa 1 para a etapa 2, podendo atingir bons resultados, mas seu limite inferior é baixo, podendo ter um baixo desempenho. Além disso, é também instável. KNN teve um desempenho ineficiente na primeira fase e continuou nas outras fases. O método SVM teve um comportamento muito parecido com o NB, conseguiu melhorar seu desempenho na segunda etapa, pode atingir bons resultados, mas é um método instável. Resumindo, para este poço, na série de treinamento, o desempenho dos métodos não foi eficiente. Mas dentre eles, o método que obteve melhor desempenho (considerando a média do AUC obtido) foi o SVM, na segunda etapa com a seleção dos parâmetros.

Após analisados os resultados obtidos na série de treinamento, as previsões obtidas na série de validação são avaliadas. Relembrando, estes modelos calibrados (de acordo com etapa 1, etapa 2 e etapa 3) pelo conjunto de treinamento são utilizados para

prever os resultados da série de validação. Os resultados previstos para cada uma das etapas são mostrados na Figura 33.

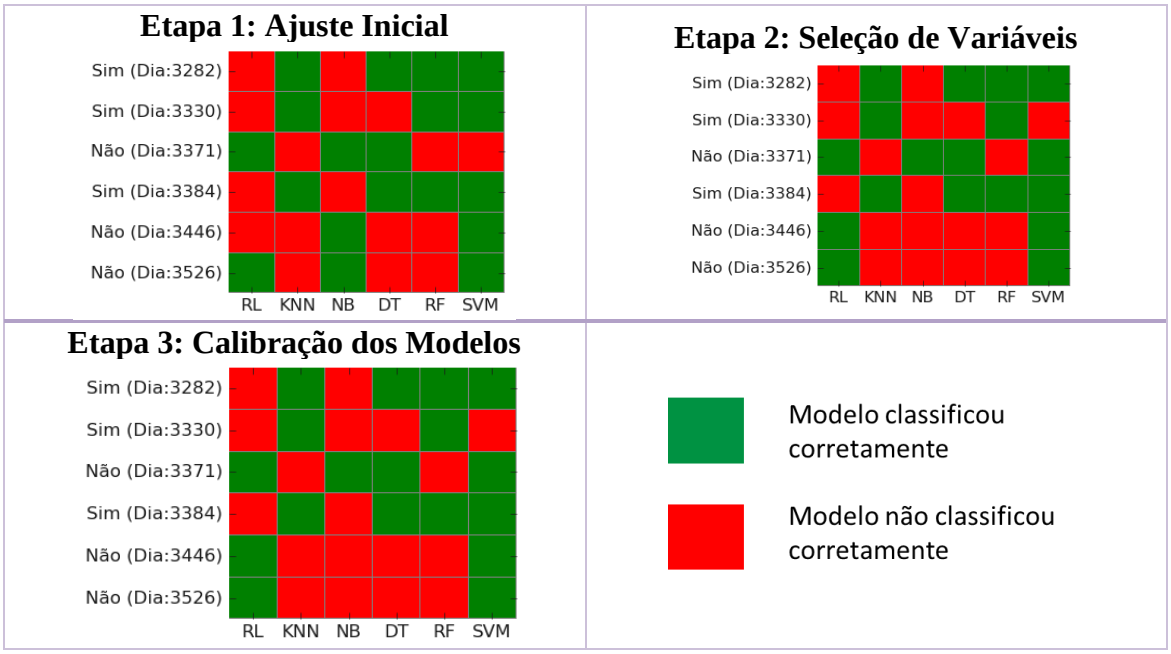


Figura 33: Resultados obtidos na série de validação para cada um dos métodos nas três etapas analisadas. Poço W1.

Conforme pode ser visto na Figura 33, cada linha corresponde a um determinado teste de produção da série de validação a ser previsto. Cada coluna é um determinado modelo de classificação, e as cores são de acordo com os resultados que os modelos conseguiram ou não prever corretamente. Na descrição da linha tem o dia que o teste foi feito para o poço analisado com sua classificação real. Por exemplo, a primeira linha corresponde a um teste de produção categorizado como válido, o terceiro teste já é um teste de produção inválido. O objetivo dos modelos é conseguir prever corretamente os testes de produção. Assim, quanto mais quadrados verdes o modelo obtiver, melhor seu desempenho. Uma outra questão é sua capacidade de prever testes do tipo *Não*. Pelo desbalanceamento da série, se o modelo de classificação não estiver corretamente calibrado e otimizado, ele não vai conseguir prever testes tipo *Não*. Dessa forma, como a série de validação contém 3 testes tipo *Sim* e 3 testes tipos *Não*, e o modelo não conseguir identificar nenhum teste tipo *Não*, o modelo é muito insatisfatório. Entretanto, é igualmente não aceitável obter um modelo que só categorize testes como sendo tipo *Não*.

A ideia principal é que para a taxa de erro do modelo, é preferível um determinado teste ser classificado como tipo *Não* (teste de produção inválido), mesmo

sendo tipo *Sim* (teste de produção válido), mas este erro não pode ser grande suficiente de modo que os testes analisados estejam errados e classificados apenas em uma categoria. Esta ideia pode ser exemplificada pelos resultados obtidos Figura 33 pelo poço W1. Analisando primeiramente o resultado do SVM, que foi o modelo com melhor resultado da série de validação. Na etapa 1, o SVR consegue prever 5 dos 6 testes, errando um teste de produção tipo *Não*. Com a seleção das variáveis, o modelo continua errando um teste, mas agora da categoria *Sim*. Para este estudo, a Etapa 2 tem um desempenho melhor do que a Etapa 1. Isto porque, apesar dos dois casos acertarem igualmente 5 dos 6 testes de produção, o modelo na etapa 2 está conseguindo prever todos os testes tipo *Não*, e está errando apenas um teste tipo *Sim*. Em situações práticas, os operadores ficariam atentos com a resposta, mas seria apenas um falso alarme. No contrário, quando um teste de produção é considerado válido sem ser, providências podem não ser tomadas para verificar a situação. Para o poço W1, o método SVR foi o único com bom desempenho. Os demais não são satisfatórios. Para escolher uma etapa, seria selecionada a etapa 2, já que não houve melhorias da etapa 2 para etapa 3.

Os gráficos das análises individuais dos demais poços para cada uma das etapas, nas séries de treinamento e validação, conforme mostradas na Figura 31, Figura 32 e Figura 33, estão em APÊNDICE I – RESULTADOS DA ETAPA DE CLASSIFICAÇÃO. O resumo dos resultados de todos os poços é mostrado nas Tabela 14 e Tabela 15. Tabela 14 contém os resultados os poços W1 a W7 e Tabela 15 dos poços W8 até W13.

Tabela 14: Resultado obtidos na etapa de classificação para os poços de W1 a W7.

POÇO	MODELO	ETAPAS						MELHOR MODELO	
		TREINAMENTO			VALIDAÇÃO			TREINO	TESTE
		1	2	3	1	2	3		
		$\overline{AUC_1}$	$\overline{AUC_2}$	$\overline{AUC_3}$	$\overline{AUC_1}$	$\overline{AUC_2}$	$\overline{AUC_3}$		
W1	RL	0,32	0,53	0,61	0,33	0,50	0,50	SVM	SVM
	KNN	0,50	0,50	0,50	0,50	0,50	0,50		
	NB	0,44	0,62	0,62	0,50	0,17	0,17		
	RF	0,57	0,46	0,46	0,50	0,50	0,50		
	SVM	0,42	**0,66**	0,64	**0,83**	**0,83**	**0,83**		
W2	RL	0,50	**0,69**	**0,69**	**1,00**	**1,00**	**1,00**	RL	RL
	KNN	0,53	0,59	0,55	0,50	0,50	0,50		
	NB	0,52	0,58	0,58	0,67	0,50	0,50		
	RF	0,65	0,67	0,49	0,50	0,67	0,50		
	SVM	0,53	0,62	0,64	0,50	0,50	0,50		
W3	RL	0,61	**0,71**	0,60	0,50	0,50	0,50	RL/NB	NB/ SVM
	KNN	0,65	0,65	0,50	0,50	0,50	0,50		
	NB	0,59	**0,70**	0,70	**0,67**	**0,67**	**0,67**		
	RF	0,54	0,55	0,60	0,50	0,50	0,50		
	SVM	0,60	0,64	0,50	0,50	**0,67**	0,50		
W4	RL	0,54	0,56	0,55	**0,83**	**0,83**	0,67	RF	RL
	KNN	0,48	0,53	0,49	0,50	0,50	0,50		
	NB	0,62	0,64	0,64	0,67	0,67	0,67		
	RF	**0,67**	0,67	0,62	0,50	0,50	0,50		
	SVM	0,60	0,64	0,57	0,67	0,67	0,67		
W5	RL	0,44	0,62	0,60	0,50	0,50	0,50	NB	SVM
	KNN	0,60	0,60	0,50	0,50	0,50	0,50		
	NB	0,65	**0,68**	0,68	0,50	0,50	0,50		
	RF	0,59	0,58	0,60	0,50	0,50	0,50		
	SVM	0,45	0,62	0,50	0,50	**0,67**	0,50		
W6	RL	0,75	**0,81**	0,80	**0,67**	**0,67**	**0,67**	RL/NB	RL/NB
	KNN	0,53	0,60	0,55	0,50	0,50	0,50		
	NB	0,74	**0,78**	0,78	0,50	**0,67**	**0,67**		
	RF	0,68	0,58	0,53	0,50	0,50	0,50		
	SVM	0,62	0,73	0,50	0,50	0,50	0,50		
W7	RL	0,72	**0,79**	0,79	0,50	0,67	0,67	SVM/RL	SVM
	KNN	0,55	0,58	0,50	0,50	0,50	0,50		
	NB	0,68	0,76	0,76	0,50	0,50	0,50		
	RF	0,62	0,68	0,68	0,50	0,50	0,50		
	SVM	0,78	**0,79**	0,50	**1,00**	**1,00**	0,50		

Tabela 15: Resultado obtidos na etapa de classificação para os poços de W8 a W13.

POÇO	MODELO	ETAPAS						MELHOR MODELO	
		TREINAMENTO			VALIDAÇÃO			TREINO	TESTE
		1	2	3	1	2	3		
		$\overline{AUC_1}$	$\overline{AUC_2}$	$\overline{AUC_3}$	$\overline{AUC_1}$	$\overline{AUC_2}$	$\overline{AUC_3}$		
W8	RL	0,67	0,75	0,75	0,67	0,67	0,50	SVM/RL	SVM
	KNN	0,56	0,69	0,54	0,50	0,67	0,50		
	NB	0,55	0,58	0,58	0,50	0,50	0,50		
	RF	0,59	0,59	0,56	0,67	0,50	0,67		
	SVM	0,65	**0,76**	0,69	**0,83**	**0,83**	0,50		
W9	RL	0,46	0,49	0,46	0,33	0,33	0,33	SVM	SVM
	KNN	0,59	0,64	0,50	0,50	0,67	0,50		
	NB	0,63	0,71	0,71	0,67	0,67	0,67		
	RF	0,68	0,68	0,66	0,67	0,67	0,67		
	SVM	0,63	**0,72**	0,50	**0,83**	0,67	0,50		
W10	RL	0,59	0,60	0,58	0,33	0,33	**0,83**	KNN	RF/RL
	KNN	0,61	**0,62**	0,62	0,50	0,50	0,50		
	NB	0,51	0,60	0,60	0,50	0,50	0,50		
	RF	0,59	0,55	0,55	0,67	**0,83**	0,67		
	SVM	0,50	0,56	0,58	0,67	0,67	0,67		
W11	RL	0,77	**0,83**	0,76	0,67	**0,83**	0,67	SVM/RL	RL
	KNN	0,54	0,65	0,50	0,50	0,50	0,50		
	NB	0,78	0,78	0,78	0,50	0,50	0,50		
	RF	0,74	0,77	0,73	0,50	0,50	0,50		
	SVM	0,83	**0,84**	0,50	0,50	0,50	0,50		
W12	RL	0,64	0,71	0,69	**0,67**	0,50	0,50	SVM	SVM/ RL/ RF
	KNN	0,56	0,60	0,53	0,50	0,50	0,50		
	NB	0,65	0,67	0,67	0,50	0,50	0,50		
	RF	0,60	0,57	0,54	0,50	**0,67**	0,50		
	SVM	0,65	0,71	**0,72**	0,50	0,50	**0,67**		
W13	RL	0,57	0,66	0,72	0,67	0,50	0,50	SVM/NB	KNN/ NB/ SVM
	KNN	0,68	0,68	0,50	**0,83**	0,67	0,50		
	NB	0,77	0,78	0,78	0,67	**0,83**	**0,83**		
	RF	0,65	0,50	0,59	0,50	0,67	0,67		
	SVM	0,76	**0,80**	**0,80**	**0,83**	**0,83**	0,67		

Sintetizando os resultados obtidos na Tabela 14 e Tabela 15, nove poços dos trezes analisados conseguiram obter bons ou excelentes resultados na série de validação, considerando uma faixa de corte de AUC de 0,83, o que corresponde um acerto de pelo menos 5 dos 6 testes disponíveis. Além disso, em 10 poços, o melhor modelo da série de treinamento é o mesmo da série de validação, o que mostra consistência nos resultados obtidos. Em situações ideias, espera-se que o melhor resultado na série de validação seja do melhor modelo selecionado pela série de treinamento. Outro ponto

importante é que na série de validação, considerando que alguns poços obtiveram mais de um melhor modelo de classificação, ou seja, modelos com mesma medida média de AUC, podem ser analisados 18 casos no total (desconsiderando os modelos de RF pela falta de consistência dos resultados). Destes 18 melhores modelos selecionados na série de validação, em 15 casos o resultado manteve-se ou foi melhorado com a retirada das variáveis (etapa 2). Ou seja, para estes 15 casos, a seleção de variáveis impactou o resultado positivamente ou removeu variáveis não necessárias para a análise de classificação. Entretanto, em relação a terceira etapa, observa-se através da análise das tabelas, que apenas em 2 dos 18 casos, os resultados melhoraram ao se aplicar a etapa 3. Os resultados na terceira etapa, que supostamente tinha a função de melhorar os modelos, ou não fizeram efeito ou pioraram o resultado. Na série de treinamento também se percebe uma piora dos resultados de AUC para muitos modelos na terceira etapa. O que indica que esta terceira etapa poderia ser desconsiderada para o estudo proposto.

Os resultados obtidos pela etapa de classificação dos testes de produção em poços de petróleo foram satisfatórios, em vista da dificuldade do problema de validação dos testes de produção, mas algumas limitações foram identificadas. Observou-se muita variabilidade das métricas de AUC em cada modelo, sendo muito dependentes das amostras selecionadas para cada ciclo de validação. Entretanto, observa-se uma potencial capacidade de melhoria dos modelos. Regressão logística e Máquina de Vetores de Suporte foram os dois métodos com melhores resultados, de forma que a atenção para melhorar os resultados pode ser dada a esses dois modelos. Além disso, é necessário revisar e melhorar o método para seleção das variáveis, já que em três casos a seleção piorou o desempenho.

Em relação a análise das variáveis na etapa 2, os resultados obtidos podem ser vistos na Figura 34. Na figura, as colunas correspondem aos poços analisados. Para cada poço, a primeira linha refere-se ao método que apresentou melhor resultado na série de validação, de acordo com o que foi mostrado na Tabela 14 e na Tabela 15. É possível notar que os poços W3, W6, W12 e W13 contêm mais de uma coluna. Estes poços são os casos em que mais de um método de classificação foi selecionado pela análise. Além do método, é mostrada a medida média de AUC obtida na série de validação, para cada um dos métodos selecionados. Após isso, a matriz com as variáveis selecionadas é indicada. Nos casos em que o método não selecionou a variável, ela toma

valor zero, e caso tenha sido selecionada, é adotado o valor 1. Além disso, caso a variável não tenha sido considerada no estudo original, o espaço fica vazio. É o caso da P1 e $\Delta P1$. Mesmo para os poços com poucos dados ausentes de P1, a variável não foi considerada na etapa de classificação, pelo baixo desempenho obtido ao se selecionar esta variável no estudo. Os dados anômalos são removidos apenas no conjunto dos testes de produção considerados válidos. A medida de P1, por ser muito inconsistente, ao ser considerada na etapa de classificação criava modelos tendenciosos na série de treinamento, colocando a invalidade dos testes somente nesta variável, o que não é uma verdade, conforme foi observado na etapa de pré-processamento. Ao não se considerar as variáveis P1 e $\Delta P1$, os modelos de classificação obtidos ficaram mais robustos e consistentes.

Ainda em relação a Figura 34, é possível notar que as variáveis selecionadas por cada método podem ser diferentes. Além disso, cada poço, por ter um comportamento diferente, irá selecionar o conjunto de variáveis próprio para seu caso. No entanto, algumas variáveis foram mais selecionadas, conforme pode ser visto na última coluna, que representa o número de vezes que a variável foi escolhida. As variáveis RGLI, vazão de óleo e $\Delta P2$ foram as variáveis selecionadas com maior recorrência pelos modelos, podendo indicar sua importância na etapa de classificação dos testes de produção.

Poço	W1	W2	W3	W3	W4	W5	W6	W6	W7	W8	W9	W10	W11	W12	W12	W13	W13	W13	
Método	SVM	RL	NB	SVM	RL	SVM	RL	NB	SVM	SVM	SVM	RL	RL	RL	SVM	KNN	NB	SVM	
AUC	0,83	1,00	0,67	0,67	0,83	0,67	0,67	0,67	1,00	0,83	0,83	0,83	0,83	0,67	0,67	0,83	0,83	0,83	SOMA
Dias	0	0	0	0	1	0	0	0	1	1	1	1	1	1	1	1	0	1	10
P1																			0
P2	0	1	0	1	0	0	1	1	1	1	1	1	0	1	1	1	0	1	**12**
P3	0	0	1	1	1	0	0	1	0	0	1	1	1	1	0	1	0	0	9
P4	0	0	0	0	1	1	0	1	1	0	1	1	0	1	0	1	1	1	10
T1	1	0	0	0		1							1						3
T2	0	1	0	0	1	0	1	0	1	1	1	1	0	1	1	1	0	1	11
T3	1	1	0	0	0	1	0	0	1	1	1	0	0	1	0	1	0	0	8
Qbruta	1	0	0	0	1	0	1	1	1	0	1	1	0	1	1	1	0	1	11
Qgl	0	1	0	0	1	0	1	0	1	1	1	1	1	1	1	1	0	0	11
Pgl	0	0	0	0	1	1	0	0	1	1	1	1	1	1	1	1	1	1	12
FE	0	0			1	0	0	1	1	1	1	1		1	1	1	0	0	9
RS	1	1			1	0	0	0	1	1	1	1		1	1	1	0	0	10
RGO	1	1	1	0	1	0	0	1	1		1	1	0	1	1	1	1	0	**12**
RGLI	1	1	1	0	1	0	1	1	1	1	1	1	1	1	1	1	0	1	**15**
BSW	1	1	0	0	1	0	0	0	1	0	1	1	1	1	1	1	0	0	10
deltaP1																			0
deltaP2	1	0	1	0	1	1	1	1	1	0	1	1	1	1	1	1	0	0	**13**
Qoleo	1	0	0	1	1	0	1	1	1	0	1	1	1	1	1	1	0	1	**13**
Qagua	0	1	0	0	1	0	0	0	1	1	1	1	1	1	0	1	0	0	9
Qgp	1	0	0	1	1	1	0	0	1		1	1	0	1	1	1	1	1	**12**
Qgt	1	0	0	0	1	0	0	0	1	0	1	1	0	1	1	1	0	1	9

Não Alterou
Melhorou com Etapa2
Melhorou com Etapa 3
Piorou com Etapa 2

0 Variável não selecionada
1 Variável selecionada
Variável não utilizada

Figura 34: Resultado das variáveis selecionadas na etapa de classificação.

6.3 Previsão das Variáveis do Processo

Nesta etapa foram estudados os modelos de regressão MLR, SVR, RT, RFR, conforme visto na metodologia apresentada no item 5.4. Além disto, foi proposta a elaboração de um intervalo de predição utilizando o método *bootstrap*. O objetivo era prever as vazões de óleo, água e gás da série de validação, que são as quatro últimas datas dos registros disponibilizados, a partir das outras variáveis analisadas. As fases de ajuste inicial dos parâmetros, seleção de variáveis e otimização do modelo SVR na série de treinamento foram conduzidas. Diferentemente dos resultados mostrados na etapa de classificação, nesta parte a análise será feita somente nas datas previstas pelo modelo para série de validação, em vista dos excelentes resultados obtidos pelas medidas de erro para todos os modelos na série de treinamento.

Após ajustar os modelos utilizando a série de treinamento, as 10000 amostras da variável dependente eram geradas pelo procedimento mostrado na metodologia. Desta análise, percebeu-se que os métodos RT e RFR exigiam um tempo computacional muito elevado e uma qualidade muito baixa nos resultados. Dessa forma, os métodos RT e RFR foram excluídos da análise. Em compensação, os métodos MLR e SVR obtiveram um desempenho muito satisfatório nas previsões das vazões de óleo e de água dos poços.

Para avaliar os resultados obtidos, foram gerados histogramas como o mostrado na Figura 35. O histograma contém todos os resultados de vazão das 10000 rodadas geradas pelo método *bootstrap* para a data prevista. A partir das amostras geradas, são determinados os percentis 10% e 90% para criar o intervalo de previsão considerado válido para análise. Este intervalo entre 10% e 90%, corresponde a 80% dos dados centrais gerados pelo método de amostragem. Assim, todos os registros que caírem dentro deste intervalo são considerados consistentes, porém registros que obtiverem resultados fora deste intervalo requerem uma atenção. Não necessariamente significa que o teste de produção é inválido, pois as vazões modificam de comportamento ao longo do tempo. Pode ser o caso de um resultado obtido fora do intervalo. De qualquer forma, para os dois motivos, mudança de comportamento e validação de testes de produção de petróleo, este comportamento requer a atenção dos operadores. Outro

aspecto importante é visualizar o quão distante um determinado valor de vazão está da distribuição de vazões esperadas, para os casos que a vazão está fora do intervalo.

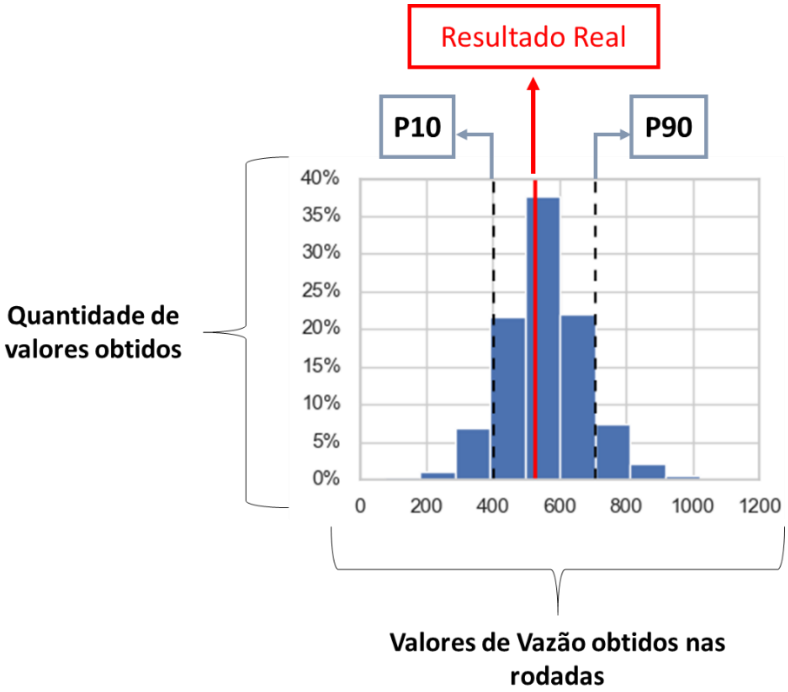


Figura 35: Esquema dos gráficos gerados para os resultados obtidos na regressão.

Na análise dos histogramas obtidos para as vazões de óleo, água e gás, procurou-se também identificar se os resultados obtidos estavam perto do P50, que é a mediana do conjunto amostrado. Quanto mais próximo do P50, mais um indicativo de que o modelo de regressão obteve um bom desempenho.

Os resultados obtidos para todos os poços, nas quatro datas previstas da série de validação são mostrados nas tabelas mostradas a seguir. A Tabela 16 e a Tabela 17 são os resultados de vazão de óleo previstos para os poços W1 até W7, e W8 até W13 respectivamente. A Tabela 18 e Tabela 19 mostram os resultados para as vazões de água. A Tabela 20 e a Tabela 21 indicam os resultados para as vazões de gás total. Nestas tabelas estão as informações do nome do poço analisado, o melhor modelo de regressão de acordo com a menor medida de RMSE obtida, as 4 datas que o teste foi realizado, e para cada uma delas a medida de erro RMSE encontrada no ajuste do modelo da série de treinamento original, as medidas de vazão para P10, P50 e P90, o valor de vazão real e o valor do ajuste obtido pelo modelo de regressão (Yaj). Além disso, para cada teste é indicado se a vazão estava dentro do intervalo de previsão

esperado. As linhas marcadas em vermelho são os casos em que os dados reais estavam fora do intervalo.

Tabela 16: Resultado obtidos na previsão de vazão de óleo para os poços de W1 a W7.

Poço	Modelo	Dia	RMSE	P10	P50	P90	Qóleo	Yaj	Previsão
W1	SVR	3282	75	291,8	451,2	626,2	454,8	687,4	✓
		3330	75	534,8	682,9	836,1	669,3	708,8	✓
		3384	74	402,9	551,0	708,4	524,8	651,1	✓
		3404	74	417,7	547,9	684,1	599,3	578,0	✓
W2	SVR	3265	17	327,6	341,5	394,6	332,7	360,4	✓
		3305	17	370,3	385,0	424,1	385,2	392,1	✓
		3343	17	349,2	361,2	397,3	357,8	378,1	✓
		3382	17	365,3	378,2	410,0	378,3	378,6	✓
W3	MLR	3478	27	667,0	704,5	739,8	697,7	714,0	✓
		3557	27	668,6	696,7	726,9	688,7	701,5	✓
		3597	27	640,7	664,7	695,4	666,5	668,4	✓
		3635	26	631,0	655,7	685,9	653,9	660,7	✓
W4	MLR	3281	4	436,1	439,1	441,7	437,5	438,5	✓
		3315	4	446,7	449,5	452,0	446,9	449,0	✓
		3393	4	416,0	418,6	421,1	419,4	418,1	✓
		3479	4	478,6	481,3	483,7	474,8	480,7	×
W5	MLR	3427	101	358,6	433,2	491,3	403,6	435,6	✓
		3491	100	273,5	346,4	414,1	348,1	352,9	✓
		3530	99	346,0	417,9	493,2	418,9	425,5	✓
		3560	98	337,2	412,4	480,3	450,9	418,6	✓
W6	SVR	3556	42	761,0	805,1	841,4	798,6	858,8	✓
		3595	42	762,2	807,4	838,9	823,6	854,0	✓
		3636	42	636,8	680,1	713,3	690,0	787,6	✓
		3656	41	643,4	683,9	715,2	688,4	761,3	✓
W7	SVR	3461	31	441,1	497,1	577,0	466,7	644,8	✓
		3498	31	493,7	542,1	618,3	522,7	574,2	✓
		3536	30	494,3	541,5	616,3	509,8	524,1	✓
		3567	30	493,9	537,6	611,5	521,4	521,6	✓

Tabela 17: Resultado obtidos na previsão de vazão de óleo para os poços de W8 a W13.

Poço	Modelo	Dia	RMSE	P10	P50	P90	Qóleo	Yaj	Previsão
W8	SVR	3513	19	187,1	206,8	221,9	220,9	249,4	✓
		3558	18	206,6	227,6	243,0	235,9	279,0	✓
		3593	18	227,5	243,1	257,0	250,9	256,5	✓
		3631	18	192,3	209,9	224,6	214,2	247,5	✓
W9	SVR	3325	84	285,3	362,6	454,9	456,0	536,6	×
		3440	84	265,8	348,7	441,8	429,2	497,0	✓
		3480	84	276,0	369,3	459,3	462,5	644,8	×
		3522	82	237,5	335,0	422,5	424,2	545,0	×
W10	SVR	3320	123	360,2	412,7	479,7	435,8	457,9	✓
		3359	122	357,4	411,8	467,0	452,0	474,0	✓
		3399	122	367,6	423,8	482,0	447,5	483,6	✓
		3438	122	382,2	437,8	494,6	449,1	478,2	✓
W11	SVR	3478	73	296,3	322,0	364,4	317,1	430,4	✓
		3518	73	288,2	313,2	355,2	309,8	425,0	✓
		3562	72	333,3	356,8	397,2	349,7	378,6	✓
		3601	71	427,3	450,7	488,5	450,6	423,0	✓
W12	SVR	3451	110	225,1	493,6	640,9	602,0	725,5	✓
		3491	111	286,0	528,2	672,5	626,2	753,6	✓
		3529	112	238,5	480,7	624,2	528,2	997,0	✓
		3539	110	282,9	518,6	643,3	571,9	575,1	✓
W13	MLR	2805	32	280,6	310,0	350,0	317,1	313,3	✓
		2841	32	330,2	361,3	406,2	326,9	364,3	×
		2874	32	304,3	332,3	366,5	314,2	334,7	✓
		2921	32	307,8	332,6	368,2	317,5	336,2	✓

Sumarizando os resultados obtidos na Tabela 16 e Tabela 17, na etapa de previsão da vazão de óleo na série de validação, 13 poços foram analisados (W1 a W13), e para cada poço, a série de validação era composta de quatro testes de produção, de forma que no total se tinha 52 testes de produção para serem analisados. Destes 52 testes, apenas 5 testes obtiveram medidas de vazão de óleo fora do intervalo de predição. Além disso, para a maior parte destas previsões fora do intervalo, o valor da vazão de óleo está muito próximo de um dos limites do intervalo, mostrando coerência nos resultados. Outro ponto importante observado pelas tabelas é que a técnica de amostragem para consideração da incerteza da previsão teve um papel importante na previsão da vazão de óleo. Para muitos testes, o valor de vazão de óleo ajustado pelo modelo de regressão (Yaj) foi corrigido ao se considerar a técnica de amostragem, visto que o valor real da vazão está muito próximo ao P50.

Tabela 18: Resultado obtidos na previsão de vazão de água para os poços de W1 a W7.

Poço	Modelo	Dia	RMSE	P10	P50	P90	Qágua	Yaj	Previsão
W1	SVR	3282	110	697,3	798,9	1014,7	881,0	743,6	✓
		3330	110	587,8	694,2	903,8	886,0	670,5	✓
		3384	111	704,8	811,5	1023,5	941,5	739,5	✓
		3404	112	638,4	746,2	952,9	829,1	783,2	✓
W2	SVR	3265	17	220,8	247,6	279,2	216,0	218,1	×
		3305	17	171,1	201,1	241,8	177,2	177,9	✓
		3343	16	188,3	214,4	254,5	198,1	192,0	✓
		3382	16	156,7	180,3	218,4	167,6	178,1	✓
W3	MLR	3478	14	243,5	263,8	284,4	278,7	264,5	✓
		3557	14	249,5	268,6	288,2	274,1	270,8	✓
		3597	14	296,6	312,4	331,4	305,4	314,4	✓
		3635	13	326,9	342,0	359,7	336,4	343,9	✓
W4	MLR	3281	1	54,8	56,5	57,7	57,2	56,4	✓
		3315	1	51,3	52,7	53,8	54,5	52,6	×
		3393	1	55,6	57,0	58,3	54,5	57,0	×
		3479	1	55,7	57,2	58,5	62,7	57,1	×
W5	SVR	3427	69	1922,5	2014,9	2061,7	2016,6	2010	✓
		3491	69	1986,8	2066,2	2111,7	2076,8	2023	✓
		3530	69	1987,8	2058,2	2102,3	2074,2	2005	✓
		3560	69	1943,8	2010,9	2053,3	1973,5	1969	✓
W6	SVR	3556	43	1261,4	1296,7	1339,9	1299,1	1237	✓
		3595	42	1204,2	1238,4	1279,0	1221,5	1201	✓
		3636	42	1325,8	1358,8	1401,9	1344,9	1264	✓
		3656	41	1321,2	1354,0	1395,1	1345,6	1296	✓
W7	MLR	3461	26	556,6	592,3	645,3	490,0	590,1	×
		3498	28	484,7	523,3	570,7	476,0	521,4	×
		3536	29	488,6	527,3	568,7	501,4	526,5	✓
		3567	28	453,4	489,1	527,5	475,1	488,4	✓

Tabela 19: Resultado obtidos na previsão de vazão de água para os poços de W8 a W13.

Poço	Modelo	Dia	RMSE	P10	P50	P90	Qágua	Yaj	Previsão
W8	SVR	3513	14	161,0	184,1	212,8	147,8	145,5	×
		3558	14	142,7	168,1	196,7	162,5	124,3	✓
		3593	13	93,7	117,5	143,9	112,8	131,4	✓
		3631	13	136,8	159,6	188,7	152,1	151,2	✓
W9	SVR	3325	47	943,0	991,2	1062,5	949,3	932,2	✓
		3440	46	986,6	1030,6	1104,4	994,1	930,5	✓
		3480	46	941,6	984,7	1057,3	955,3	886,7	✓
		3522	47	967,7	1008,8	1079,4	985,4	920,3	✓
W10	SVR	3320	38	1060,2	1117,6	1164,6	1109,6	1032	✓
		3359	37	1039,2	1094,2	1141,2	1046,8	1049	✓
		3399	37	1067,3	1123,3	1173,4	1095,6	1066	✓
		3438	37	1071,7	1124,4	1172,3	1100,4	1078	✓
W11	SVR	3478	62	2190,2	2220,8	2260,4	2227,7	2116	✓
		3518	62	2199,3	2228,2	2264,3	2232,4	2165	✓
		3562	63	2195,6	2224,0	2258,5	2232,1	2167	✓
		3601	63	2130,3	2158,9	2199,6	2157,9	2068	✓
W12	SVR	3451	112	2345,3	2541,9	2813,6	2391,3	2263	✓
		3491	112	2315,7	2503,5	2778,3	2359,9	2252	✓
		3529	112	2232,6	2411,2	2700,1	2269,3	1960	✓
		3539	110	2328,2	2485,5	2781,9	2396,6	2317	✓
W13	MLR	2805	19	826,5	852,0	873,2	861,4	850,9	✓
		2841	19	852,7	879,3	905,3	878,7	879,4	✓
		2874	19	889,6	913,9	933,9	930,9	914,3	✓
		2921	19	882,0	904,8	923,2	916,1	904,6	✓

Sumarizando os resultados obtidos na Tabela 18 e Tabela 19, em relação a vazão de água, dos 52 testes de produção para serem analisados, apenas 7 deles estão fora do intervalo de previsão. Além disso, assim como na vazão de óleo, o valor de vazão de água ajustado pelo modelo de regressão (Yaj) foi corrigido ao se considerar a técnica de amostragem, visto que o valor real da vazão está muito próximo ao P50.

Tabela 20: Resultado obtidos na previsão de vazão de gás total (10^5 m^3/dia) para os poços de W1 a W7.

Poço	Modelo	Dia	RMSE	P10	P50	P90	Qgás	Yaj	Previsão
W1	MLR	3282	0,3	3,3	3,7	4,2	2,2	3,7	×
		3330	0,4	2,8	3,4	4,0	2,5	3,4	×
		3384	0,4	2,9	3,5	4,1	2,4	3,5	×
		3404	0,4	2,5	3,1	3,7	2,5	3,1	✓
W2	SVR	3265	0,3	2,1	3,2	3,8	2,6	2,6	✓
		3305	0,3	2,8	3,8	4,5	2,9	3,3	✓
		3343	0,3	2,6	3,6	4,2	2,9	3,0	✓
		3382	0,3	2,5	3,5	4,2	2,8	2,9	✓
W3	SVR	3478	0,2	3,5	4,1	4,8	3,0	3,5	×
		3557	0,2	3,0	3,7	4,3	2,9	3,5	×
		3597	0,2	2,9	3,5	4,0	2,9	3,2	✓
		3635	0,2	2,8	3,3	3,8	2,9	3,0	✓
W4	SVR	3281	0,3	2,4	3,0	3,9	2,6	2,5	✓
		3315	0,3	2,1	2,6	3,3	2,4	2,3	✓
		3393	0,3	2,5	3,0	3,5	2,7	2,6	✓
		3479	0,3	2,5	2,9	3,4	2,6	2,5	✓
W5	SVR	3427	0,3	2,8	3,4	4,0	2,6	3,5	×
		3491	0,3	2,7	3,3	4,0	2,6	3,2	×
		3530	0,3	2,5	3,1	3,7	2,6	2,5	✓
		3560	0,3	2,5	3,1	3,7	2,6	2,9	✓
W6	MLR	3556	0,3	2,9	3,2	3,6	3,2	3,2	✓
		3595	0,3	2,4	2,7	3,1	2,8	2,8	✓
		3636	0,3	2,6	2,9	3,3	3,1	2,9	✓
		3656	0,3	2,7	3,0	3,3	3,2	3,1	✓
W7	SVR	3461	0,3	3,5	4,4	5,0	2,8	3,8	×
		3498	0,3	3,3	4,2	4,9	2,8	3,4	×
		3536	0,3	2,6	3,7	4,4	2,9	2,9	✓
		3567	0,3	2,8	3,8	4,4	2,9	2,9	✓

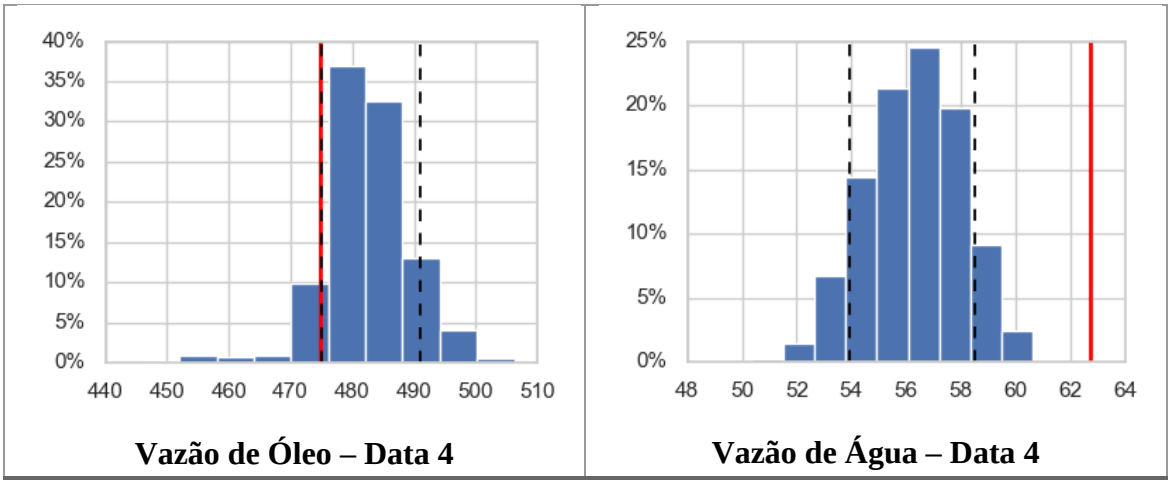
Tabela 21: Resultado obtidos na previsão de vazão de gás total (10^5 m^3/dia) para os poços de W8 a W13.

Poço	Modelo	Dia	RMSE	P10	P50	P90	Qgás	Yaj	Previsão
W8	SVR	3513	0,2	2,6	3,1	3,7	2,1	2,9	×
		3558	0,2	1,6	2,3	2,9	2,1	2,5	✓
		3593	0,2	1,9	2,5	3,0	2,0	2,5	✓
		3631	0,2	1,8	2,4	3,0	2,1	2,2	✓
W9	SVR	3325	0,3	2,3	3,4	4,0	2,4	2,5	✓
		3440	0,3	2,5	3,5	4,2	2,3	2,6	×
		3480	0,3	2,7	3,8	4,4	2,6	2,7	×
		3522	0,3	2,6	3,6	4,3	2,6	2,5	✓
W10	SVR	3320	0,2	3,0	3,9	4,6	2,7	2,9	×
		3359	0,2	2,8	3,8	4,4	2,6	2,8	×
		3399	0,2	2,8	3,8	4,3	2,7	2,7	×
		3438	0,2	2,6	3,6	4,1	2,7	2,9	✓
W11	SVR	3478	0,3	1,4	2,1	2,8	2,2	2,2	✓
		3518	0,3	1,4	2,0	2,7	2,1	2,2	✓
		3562	0,3	1,5	2,1	2,7	2,2	2,2	✓
		3601	0,3	1,7	2,3	2,9	2,3	2,2	✓
W12	SVR	3451	0,2	3,3	3,7	4,3	3,0	3,4	×
		3491	0,2	3,1	3,6	4,2	3,1	3,2	✓
		3529	0,2	3,1	3,6	4,3	3,4	3,1	✓
		3539	0,2	3,1	3,5	4,2	3,2	3,2	✓
W13	SVR	2805	0,2	2,3	3,1	3,8	2,2	2,9	×
		2841	0,1	2,4	3,1	3,7	2,3	2,7	×
		2874	0,1	2,2	2,9	3,6	2,4	2,8	✓
		2921	0,1	2,2	2,8	3,3	2,3	2,2	✓

Em relação a vazão de gás, observando as Tabela 20 e Tabela 21, dos 52 testes de produção analisados, 18 estavam fora do intervalo de previsão. Além disso, ao se observar os resultados destas tabelas, percebe-se um baixo desempenho apresentado nas análises da vazão de gás total. Esta vazão, por variar muito, cria intervalos muito instáveis.

Após análise dos resultados obtidos nas previsões das vazões de óleo e água mostrados nas tabelas Tabela 16, Tabela 17, Tabela 18 e Tabela 19, foi visto que a maior parte dos resultados foram muito bons, principalmente nas vazões de óleo. Entretanto, em algumas datas as vazões não foram previstas adequadamente, como por exemplo, o teste do dia 3479 do poço W4. O resultado dos histogramas deste poço, para este dia, é mostrado na Figura 36. Observa-se que apesar da vazão de óleo não ter sido prevista corretamente, esta está muito perto do limite inferior, o que não se mostra um resultado muito destoante. Entretanto, a vazão de água aumentou consideravelmente,

não tendo nenhuma amostra de vazão com este valor. Este comportamento gera um sinal de alerta, pois o teste ou não foi conduzido de forma correta, ou o poço está mudando de comportamento.



Vazão de Óleo – Data 4 **Vazão de Água – Data 4**

Figura 36: Resultado das vazões de óleo e água para o poço W4 no dia 3479.

Outro ponto importante é que poços com uma produção baixa de água são mais comportados, e não apresentam grande variabilidade nas vazões. Dessa forma, os modelos de regressão serão bem ajustados, e os desvios serão pequenos. Isso faz com que os valores amostrados não variem muito, estando muito próximo do valor previsto pelo ajuste da regressão. Os intervalos para estes casos são muito restritos, e qualquer variação um pouco maior pode fazer com que o valor previsto pelo modelo obtenha um resultado fora do intervalo. Até certo ponto isto é positivo, pois modelos mais regulares irão fornecer respostas mais regulares, e modelos irregulares terão respostas variando mais. Mas, dependo do caso analisado, modelos muito restritos podem constantemente obter resultados fora do intervalo, assim como modelos muito instáveis, por terem desvios muito grandes, podem aceitar qualquer resultado. Dessa forma, é sempre bom fazer a análise considerando o comportamento individual de cada poço.

Os resultados dos histogramas das vazões de óleo, água e gás, para todas as datas e todos os poços, aplicando os métodos de regressão linear múltipla e SVR, podem ser vistos em APÊNDICE II – RESULTADOS DA ETAPA DE REGRESSÃO.

Em relação a análise das variáveis selecionadas para a vazão de óleo, é possível verificar o resultado obtido pela análise da Figura 37. Cada coluna corresponde ao poço analisado, e o método de regressão considerado para cada poço é o selecionado e mostrado na Tabela 16 e na Tabela 17. Variáveis que foram selecionadas são representadas pela cor verde, registros em vermelho foram os removidos da análise, e

em amarelo foram as variáveis não consideradas na análise inicial, em função da grande quantidade de dados ausentes. A partir da análise da Figura 37, é possível verificar que as variáveis BSW e vazão bruta (Qbruta) foram sempre selecionadas, o que está de acordo com o processo físico esperado do processo. A variável Dias também foi selecionada para muitos casos, o que também é esperado, visto que a vazão de óleo pode modificar seu comportamento ao longo do tempo de produção. A variável RGLI também foi selecionada para muito dos casos, e também está de acordo com o processo físico. É possível notar que a variável P1 foi considerada para alguns dos poços analisados. Como a etapa de previsão das variáveis utilizou apenas testes de produção válidos, para alguns poços foi possível fazer um refino dos dados dessa variável. Com o pré-processamento, foi possível considerar P1 na análise dos poços W3, W5, W6 e W11, e inclusive, a variável foi selecionada para três dos quatro casos.

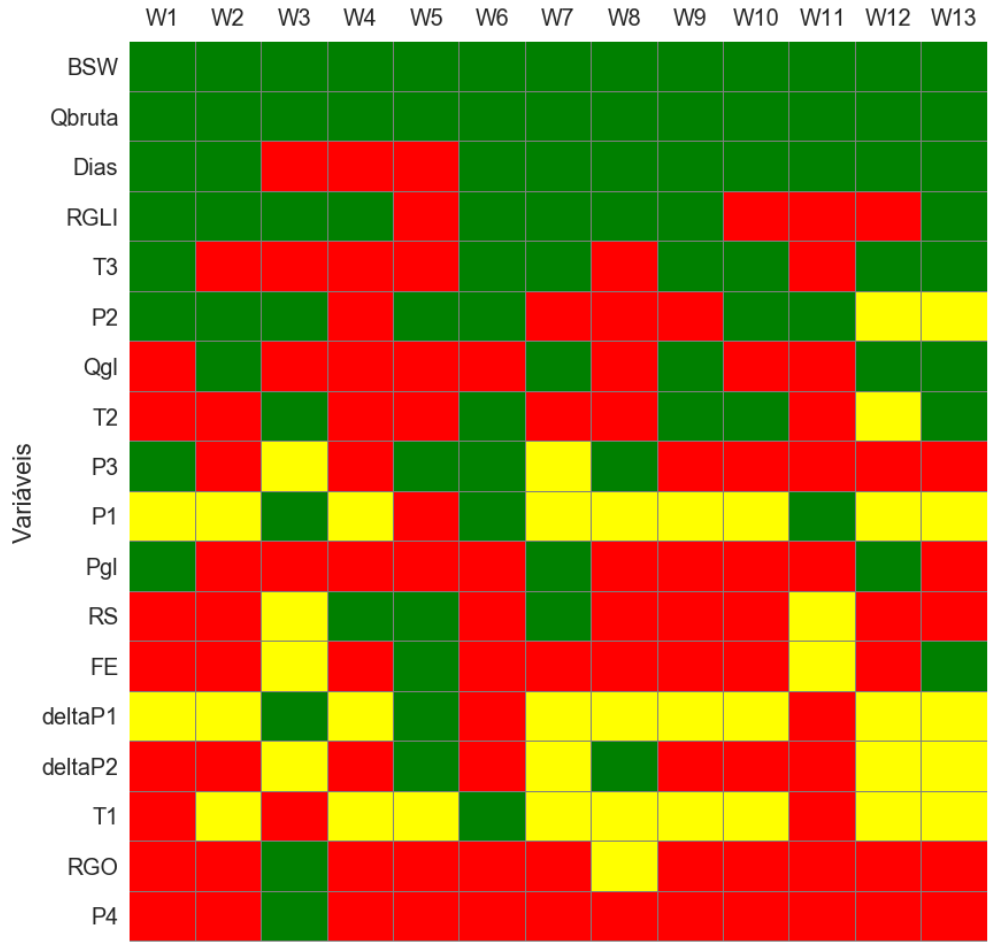


Figura 37: Resultado das variáveis selecionadas para vazão de óleo.

Ainda em relação a Figura 37, T3 é uma variável não muito recorrente nos estudos de previsão da vazão de óleo, mas que foi selecionada com certa frequência no

resultado apresentado na Figura 37. Pensando no processo físico, a temperatura do fluido no leito marinho (T2) modifica muito ao chegar na plataforma (T3), e mudanças de temperatura acarretam em modificações nas propriedades do fluido. Dessa forma, a seleção da variável T3 para vários dos poços analisados mostra que esta temperatura pode ter influência nos resultados de vazão de óleo obtidos.

Os resultados das variáveis selecionadas para a vazão de água nos poços analisados podem ser vistos na Figura 38. Para esta vazão, a variável BSW foi selecionada para todos os poços, o que também é fisicamente esperado. A vazão de água também modifica seu comportamento ao longo do tempo, e percebe-se que a variável Dias foi selecionada na maior parte dos poços. As variáveis vazão bruta, P2, T3 tiveram influência na maior parte dos poços, para os modelos de regressão analisados.

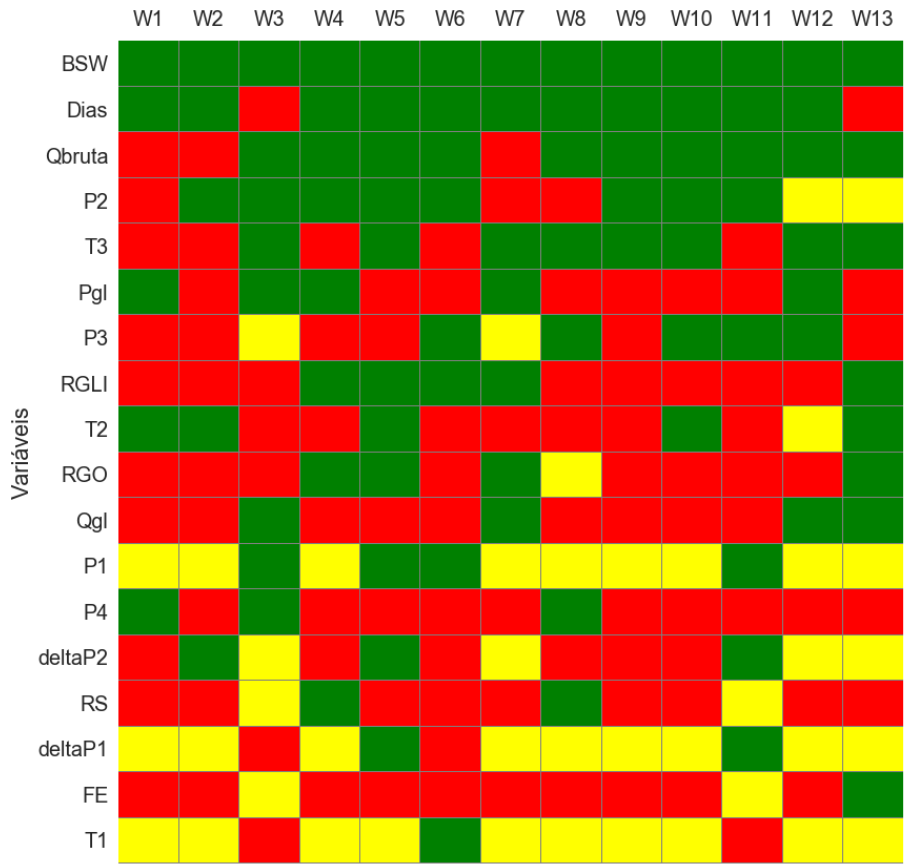


Figura 38: Resultado das variáveis selecionadas para vazão de água.

Como os resultados de vazão de gás total produzido não obteve um resultado satisfatório, a análise das variáveis selecionadas não foi feita para esta vazão.

7 CONCLUSÃO

7.1 Considerações Finais

Testes de produção em poços de petróleo têm grande importância para a atividade de produção de petróleo, visto que fornecem informações da situação corrente da produção do poço e permitem a identificação de possíveis problemas. Dada sua relevância, é necessário que os testes sejam confiáveis e de boa qualidade. Neste contexto, o estudo sobre técnicas inteligentes para validação dos testes de produção, que ainda é incipiente, precisa ser melhorado.

Este trabalho teve como objetivo desenvolver ferramentas para facilitar o processo de validação de testes de produção e colaborar com que essa validação seja feita em tempo real. Para isto, a partir de um histórico de testes de produção, modelos de mineração de dados foram estudados para classificar um teste de produção como válido ou inválido, de acordo com os dados de entrada das variáveis, e para prever vazões de óleo, água e gás, juntamente com os seus respectivos intervalos de predição. Estes intervalos foram construídos aplicando a técnica de amostragem *bootstrap*. Estas duas etapas passaram anteriormente por uma fase de pré-processamento para retirada de dados inconsistentes.

Na etapa de pré-processamento, foi percebida a importância da identificação e retirada dos dados anômalos, principalmente nas medidas de pressão que se mostraram muito inconsistentes. Da análise feita, verificou-se que mesmo testes considerados válidos podem ter medidas de pressão com resultados fisicamente impossíveis. Por isso, há uma a necessidade do refinamento do conjunto dos dados brutos para obtenção de resultados mais confiáveis. A qualidade dos resultados das etapas que seguiram o estudo dependeu muito da etapa de pré-processamento. Além disso, a utilização dos três métodos para identificar os *outliers* foi eficiente, pois permitiu a identificação de diferentes casos de existência de *outliers*. Entretanto, essas medidas servem de suporte, pois a remoção dos dados anômalos apontados pelos métodos ainda depende da análise criteriosa do usuário.

Em relação a etapa de classificação dos testes de produção, apesar da variabilidade da medida de AUC encontrada na série de treinamento, os resultados encontrados na etapa de classificação foram bem satisfatórios, quando se considera a

natureza complexa do problema estudado. Não existe nas operações normais de produção dos campos brasileiros uma ferramenta que, a partir das medições obtidas durante a realização do teste, possa classificar se o teste de produção está válido ou não. Dessa forma, o modelo de diagnóstico de testes desenvolvido que consegue fazer esta análise se mostra um grande avanço para a indústria e a atividade de produção de petróleo.

Na etapa de previsão, os resultados para a vazão de óleo e vazão de água obtiveram ótimo desempenho na maior parte dos poços com a aplicação do método de amostragem *bootstrap*. O método conseguiu incorporar a incerteza do processo, conseguindo obter bons resultados na maior parte dos casos mostrados. Um ponto interessante é que para muitos casos, os valores de vazão de óleo e água reais caíram bem próximos do P50, o que fornece uma confiabilidade maior para o resultado obtido. Com os limites obtidos pelo modelo é possível criar uma região de estabilidade, de forma que se tenha, durante a realização do teste, um controle maior das medidas que estão sendo obtidas.

Entretanto, para a vazão de gás, em função da sua alta instabilidade e altos desvios característicos da variável, o método de geração dos intervalos de previsão pode ser incipiente. Não se obteve bons resultados de previsão para muitos dos poços analisados. Neste caso, talvez seja mais eficiente a previsão de variáveis como razão gás-óleo, ou razão gás-líquido, que são frações parciais, para conseguir identificar o comportamento do gás durante um teste.

Os resultados obtidos nas etapas de classificação e regressão se mostraram bem satisfatórios. A análise de testes de produção, e de outras operações da indústria, por serem muito complexas, são muito difíceis de serem resolvidas pelos métodos convencionais de simulação. Além disso, muitas vezes, as análises dos testes de produção de petróleo são feitas de forma subjetiva, considerando apenas o conhecimento da equipe técnica responsável, o que pode estar sujeito a muitos erros. Assim, os resultados encontrados neste trabalho mostram que as técnicas de mineração de dados apresentam alta capacidade de serem aplicadas na indústria em problemas de diagnóstico, podendo trazer muitos benefícios para a atividade.

7.2 Sugestão para Trabalhos Futuros

Diferentes linhas de pesquisa podem ser sugeridas como trabalhos futuros. Em uma delas, a etapa posterior de análise horária dos testes de produção pode ser conduzida, para a identificação de como os dados se comportam durante a realização do teste. Além disso, as ferramentas sugeridas nessa dissertação poderiam ser aplicadas nesta etapa. Acredita-se que trabalhos desenvolvidos nesta linha poderão colaborar na automatização do processo de validação de testes de produção.

Outro ponto importante que é sugerido, é a análise de medidas de pressão, que se mostraram muito instáveis durante o estudo. Métodos eficientes de inserções de dados anômalos podem ser estudados para verificar o comportamento das variáveis e determinar as pressões nos registros ausentes ou de baixa confiabilidade.

Além disso, para a seleção das variáveis, heurísticas poderiam ser testadas para a determinação do subconjunto ótimo de variáveis que irá aperfeiçoar os resultados dos modelos de classificação.

Como último ponto, sugere-se a integração dos métodos de mineração de dados com os modelos dos simuladores de escoamento multifásico, para melhoria dos resultados obtidos por esses modelos dos simuladores.

8 REFERÊNCIAS BIBLIOGRÁFICAS

A. STINE, R. Bootstrap Prediction Intervals for Regression. **Journal of the American Statistical Association**, v. 80, n. 392, p. 1026–1031, 1985.

ABDELAZIZ, M.; LASTRA, R.; XIAO, J. J. ESP Data Analytics: Predicting Failures for Improved Production Performance. **Abu Dhabi International Petroleum Exhibition & Conference**, 2017.

AGGARWAL, C. C. **Data Mining: The Textbook**. [s.l: s.n.]. 2015.

ALLEN, T. O.; ROBERTS, A. P. **Production Operations: Well Completions, Workover, and Stimulation**. [s.l.] Oil & Gas Consultants International, 1982.

ARPS, J. J. Analysis of Decline Curves. **Transactions of the AIME**, v. 160, n. 01, p. 228–247, 1 dez. 1945.

BALAJI, K. et al. Status of Data-Driven Methods and their Applications in Oil and Gas Industry. 2018.

BEGGS, D. H.; BRILL, J. P. A Study of Two-Phase Flow in Inclined Pipes. **Journal of Petroleum Technology**, v. 25, n. 05, p. 607–617, 1 maio 1973.

BEGGS, H. D. **Production Optimazation Using Nodal Analysis**. Tulsa, Oklahoma: OGCI and Petroskllis Publications, 2003.

BEN-GAL, I. Outlier Detection. In: **Data Mining and Knowledge Discovery Handbook**. Boston, MA: Springer, 2005.

BRAVO, C. et al. **Applying Analytics to Production Workflows: Transforming Integrated Operations into Intelligent Operations**. SPE Intelligent Energy Conference & Exhibition. **Anais**...Society of Petroleum Engineers, 1 abr. 2014.

BREUNIG, M. M. et al. LOF. **ACM SIGMOD Record**, v. 29, n. 2, p. 93–104, 1 jun. 2000.

BRILL, J. P.; MUKHERJEE, H. **Multiphase Flow in Wells**. 1 ed. ed. Texas: [s.n.].

BRUNI, T. et al. **A Technically Rigorous and Fully Automated System for Performance Monitoring and Production Test Validation**. SPE International Improved Oil Recovery Conference in Asia Pacific. **Anais**...Society of Petroleum

Engineers, 4 abr. 2003.

CAO, Q. et al. Data Driven Production Forecasting Using Machine Learning. **SPE Argentina Exploration and Production of Unconventional Resources Symposium**, 2016.

CHAUDHARY, N. L.; LEE, W. J. Detecting and Removing Outliers in Production Data to Enhance Production Forecasting. **SPE/IAEE Hydrocarbon Economics and Evaluation Symposium**, 2016.

CRAMER, R.; JAKEMAN, S. V. J.; BERENDSCHOT, L. **Well Test Optimization and Automation**. Intelligent Energy Conference and Exhibition. **Anais**...Society of Petroleum Engineers, 4 abr. 2006.

CULLICK, A. S. et al. **A Real-Time Automated "Smart Flow" to Prioritize, Validate, and Model Production Well Testing**. SPE Digital Energy Conference. **Anais**...Society of Petroleum Engineers, 5 mar. 2013.

DUONG, A. N. An Unconventional Rate Decline Approach for Tight and Fracture-Dominated Gas Wells. **Canadian Unconventional Resources and International Petroleum Conference**, v. 90, 2010.

EFRON, B. Bootstrap Methods: Another Look at the Jackknife. **The Annals of Statistics**, v. 7, n. 1, p. 1–26, 1979.

FAYYAD, U.; PIATETSKY-SHAPIRO, G.; SMYTH, P. From Data Mining to Knowledge Discovery in Databases. **AI Magazine**, v. 17, p. 37–54, 1996.

GILBERT, W. Flowing and gas-lift well performance. **API Drilling and Production Practice**, v. 20, n. 1954, p. 126–157, 1954.

HAN, J. et al. **Data Mining Concepts and Techniques**. [s.l: s.n.]. 2012.

HUBER, P. J. Robust estimation of a location parameter. **Annals of Mathematical Statistics**, v. 35, p. 73–101, mar. 1964.

IGLEWICZ, B. AND HOAGLIN, D. The ASQC Basic References in Quality Control: Statistical Techniques. **ASQC Quality Press**, v. Vol. 16, 1993.

ILK, D. et al. Hybrid Rate-Decline Models for the Analysis of Production Performance in Unconventional Reservoirs. **SPE Annual Technical Conference and Exhibituion**, n. 1919, 2010.

JOHN, G. H.; KOHAVI, R.; PFLEGER, K. **Irrelevant Features and the Subset Selection Problem**. MACHINE LEARNING: PROCEEDINGS OF THE ELEVENTH INTERNATIONAL. **Anais**...Morgan Kaufmann, 1994

KHAN, M. R. et al. **Machine Learning Application for Oil Rate Prediction in Artificial Gas Lift Wells**. SPE Middle East Oil and Gas Show and Conference. **Anais**...Society of Petroleum Engineers, 15 mar. 2019.

KIANG, M. Y. A comparative assessment of classification methods. **Decision Support Systems**, v. 35, n. 4, p. 441–454, 2003.

KUHN, M.; JOHNSON, K. **Applied Predictive Modeling**. [s.l: s.n.]. Springer, 2013.

LYONS, W. C. **Standard Handbook of Petroleum & Natural Gas Engineering**. Vol. 2 ed. Houston, TX.: Gulf Publishing Co., 1996.

MONTGOMERY, D. C.; RUNGER, G. C. **Applied Statistics and Probability for Engineers**. Quinta edição. [s.l: s.n.].

NASER, H. H.; ZAINAL, Y. A. **Renovating Bahrain's Field: Improving Production Testing Efficiency**. SPE Middle East Oil & Gas Show and Conference. **Anais**...Society of Petroleum Engineers, 8 mar. 2015.

NAVLANI, A. **Understanding Logistic Regression in Python**. Disponível em: <https://www.datacamp.com/community/tutorials/understanding-logistic-regression-python>.

OLSEN, S.; NORDTVEDT, J.-E. **Experience from the Use of Automatic Well-Test Analysis**. SPE Annual Technical Conference and Exhibition. **Anais**...Society of Petroleum Engineers, 4 abr. 2006.

PEDREGOSA, F. et al. Scikit-learn: Machine Learning in {P}ython}. **Journal of Machine Learning Research**, v. 12, p. 2825--2830, 2011.

RAO, S. R.; DAVID, R. M. **Integrated Production Testing Framework to Improve Next Generation Production Workflows**. Abu Dhabi International Petroleum Exhibition and Conference. **Anais**...Society of Petroleum Engineers, 9 nov. 2015.

RODRIGUEZ, J. A. et al. **New Generation of Petroleum Workflow**

Automation: Philosophy and Practice. SPE Digital Energy Conference. **Anais**...Society of Petroleum Engineers, 5 mar. 2013.

ROSA, A. J.; XAVIER, J. A. D.; CARVALHO, R. DE S. **Engenharia de Reservatórios de Petróleo**. Rio de Janeiro: Editora Interciência, 2006.

ROSSI, N. C. M. **Elevação Natural de Petróleo**. Salvador: [s.n.].

SÆTEN, S. **Production Allocation of Oil and Gas: A case Study of the Skarv Field**. [s.l.] NTNU, 2015.

SUBRAHMANYA, N. et al. **Advanced Machine Learning Methods for Production Data Pattern Recognition**. SPE Intelligent Energy Conference & Exhibition. **Anais**...Society of Petroleum Engineers, 1 abr. 2014.

THOMAS, J. E. **Fundamentos de Engenharia de Petróleo**. 2 ed. ed. Rio de Janeiro: Editora Interciência, 2001.

VALKO, P. P.; LEE, W. J. A Better Way To Forecast Production From Unconventional Gas Wells. **SPE Annual Technical Conference and Exhibition**, n. September, p. 19–22, 2010.

VAPNIK, V. **The Nature of Statistical Learning Theory**. 2nd. ed. [s.l.] Springer-Verlag New York, 1995.

VELASQUEZ, G. et al. ESP "Smart Flow" Integrates Quality and Control Data for Diagnostics and Optimization in Real Time. **SPE Digital Energy Conference**, 2013.

ZANGL, G.; OBERWINKLER, C. P. **Predictive Data Mining Techniques for Production Optimization**. SPE Annual Technical Conference and Exhibition. **Anais**...Society of Petroleum Engineers, 4 abr. 2004.

APÊNDICE I – RESULTADOS DA ETAPA DE CLASSIFICAÇÃO

Neste apêndice encontram os resultados obtidos para os poços na etapa de classificação dos testes de produção. São mostrados para cada poço, os resultados da série de treinamento e os resultados da série de validação.

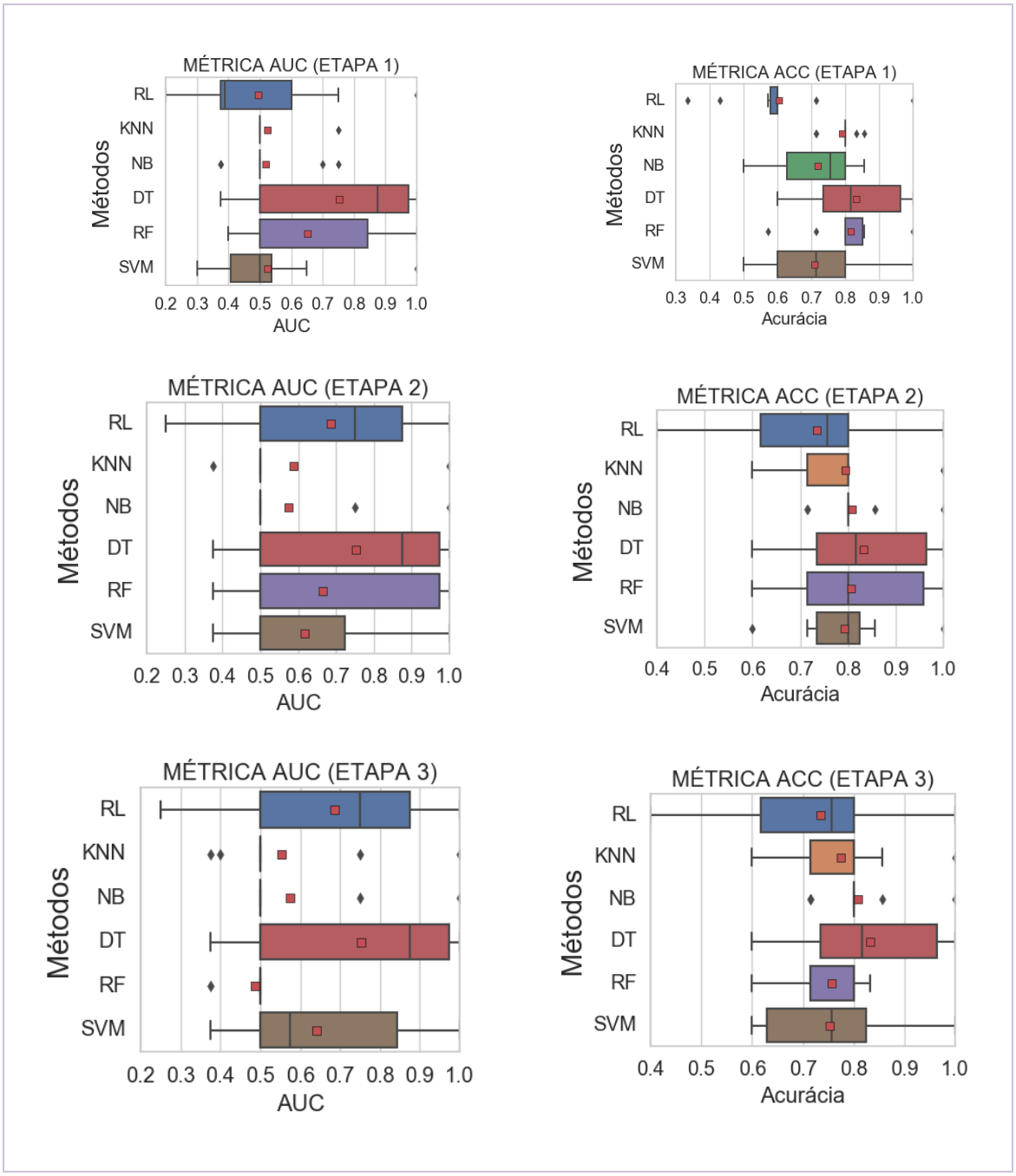


Figura 39: Métricas de AUC e acurácia (ACC) para série de treinamento do poço W2.

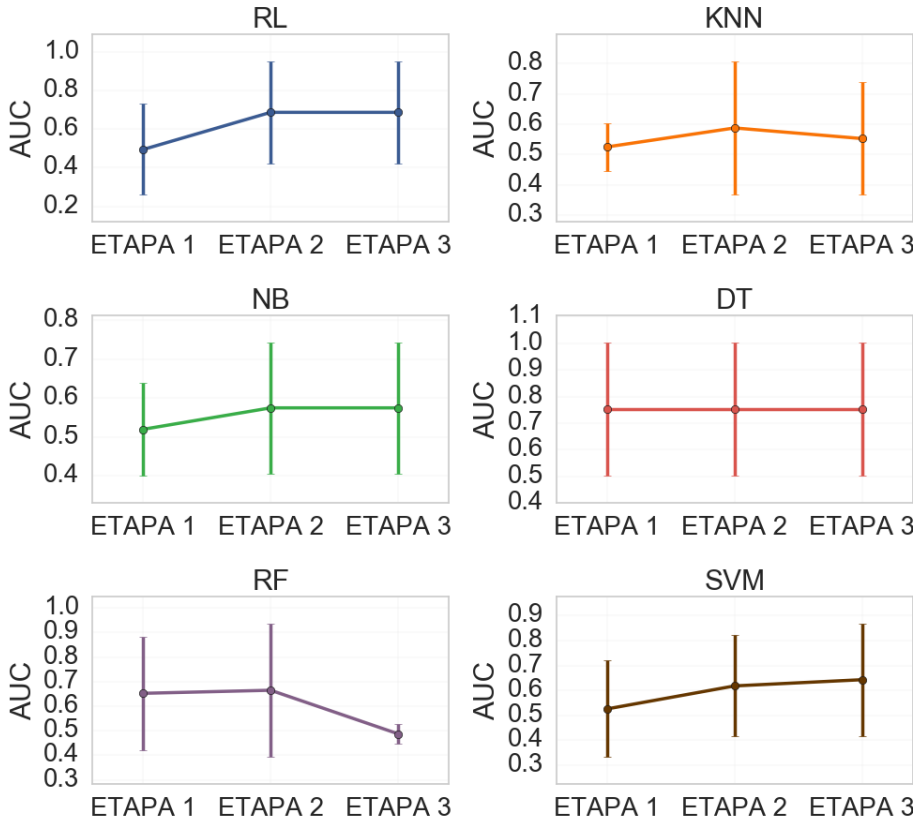


Figura 40: Médias de AUC com desvio-padrão obtidas em cada etapa da série de treinamento para poço W2.

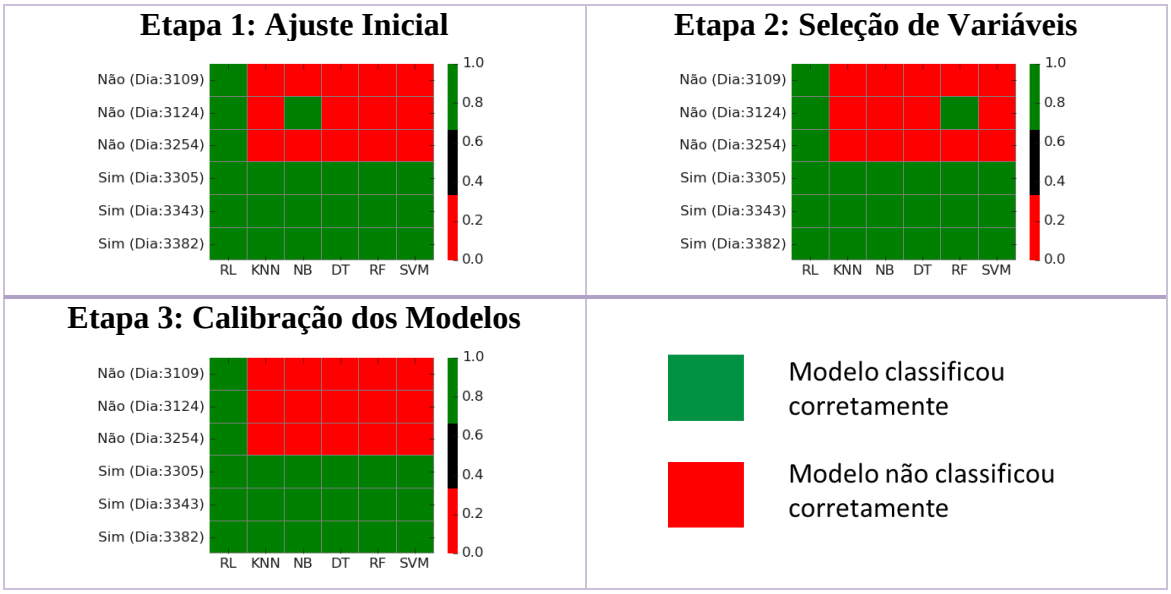


Figura 41: Resultados obtidos na série de validação para cada um dos métodos nas três etapas analisadas. Poço W2.

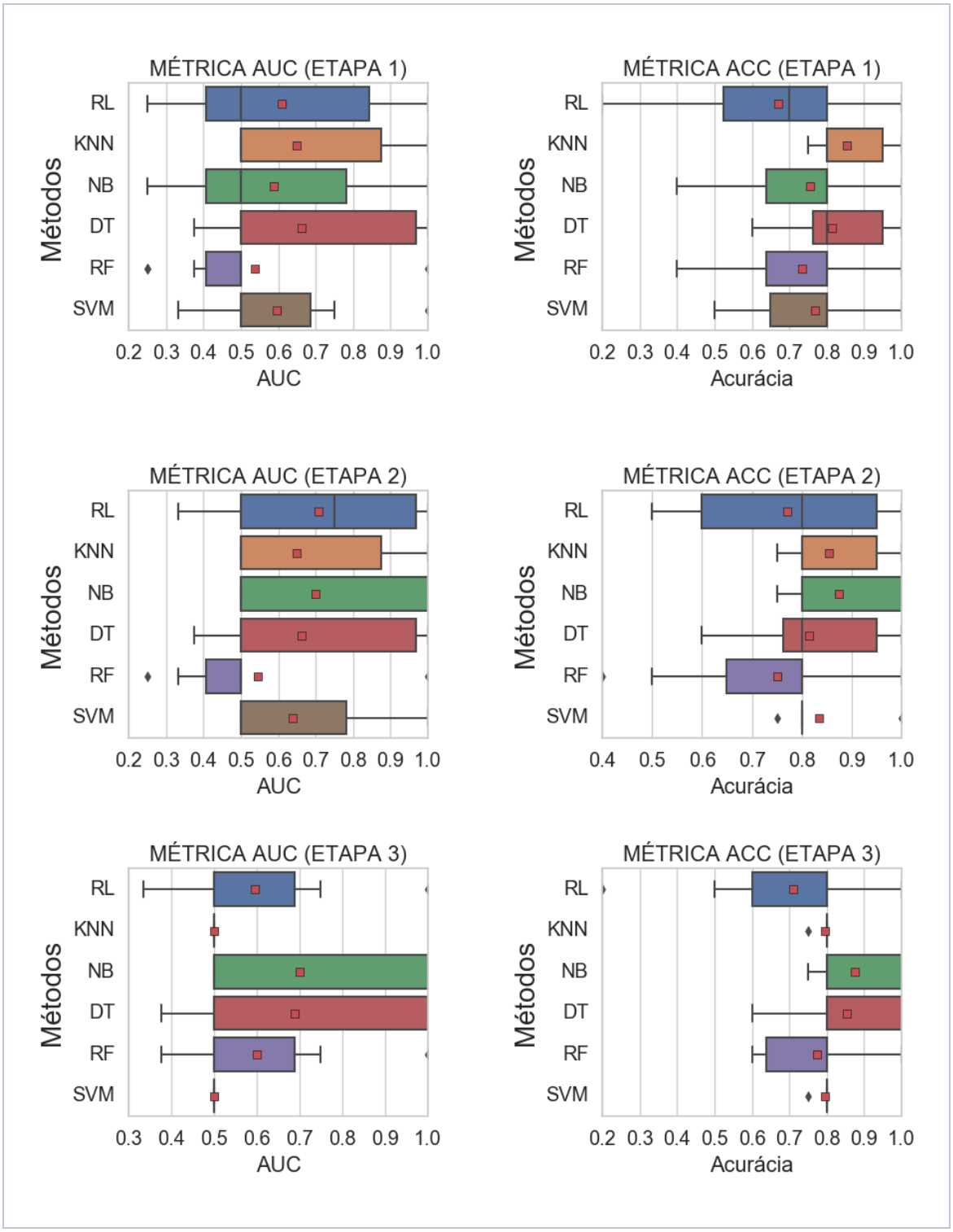


Figura 42: Métricas de AUC e acurácia (ACC) para série de treinamento do poço W3.

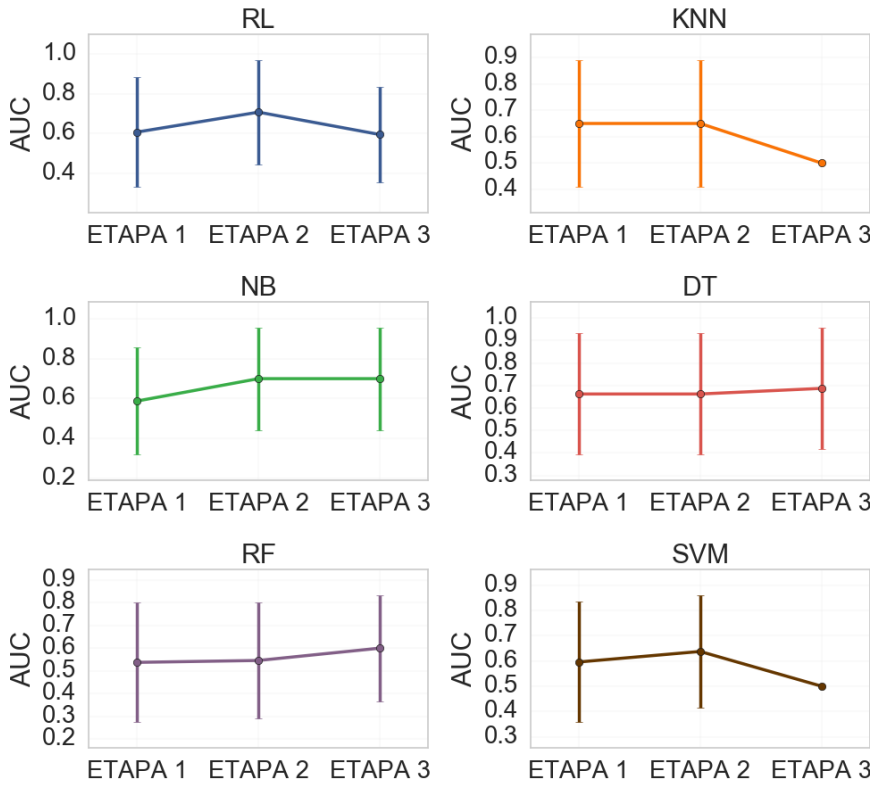


Figura 43: Médias de AUC com desvio-padrão obtidas em cada etapa da série de treinamento para poço W3.

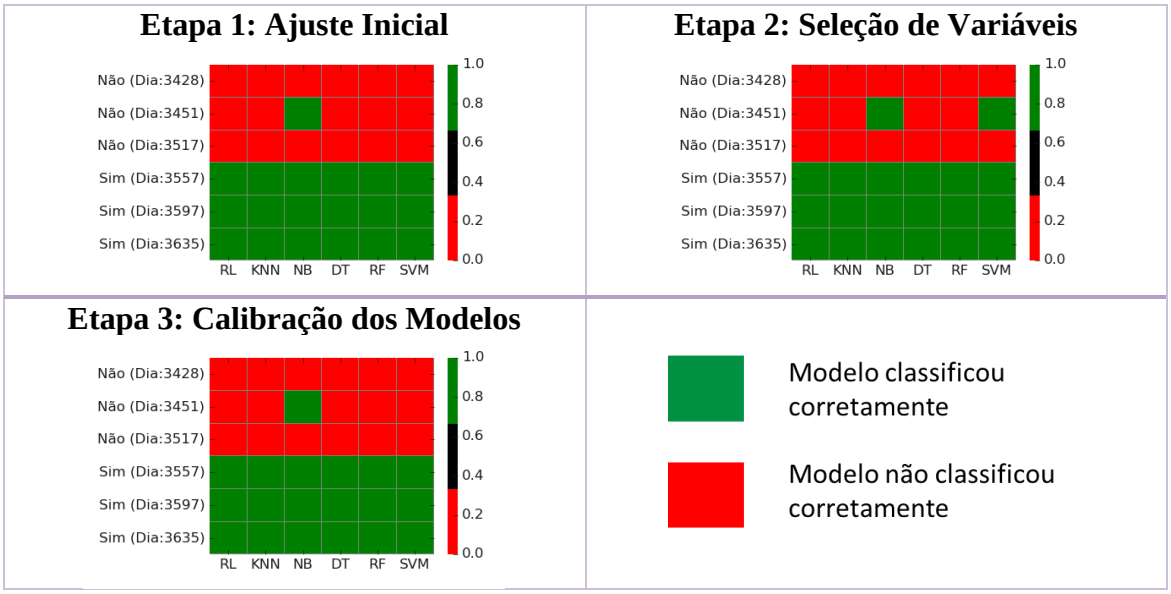


Figura 44: Resultados obtidos na série de validação para cada um dos métodos nas três etapas analisadas. Poço W3.

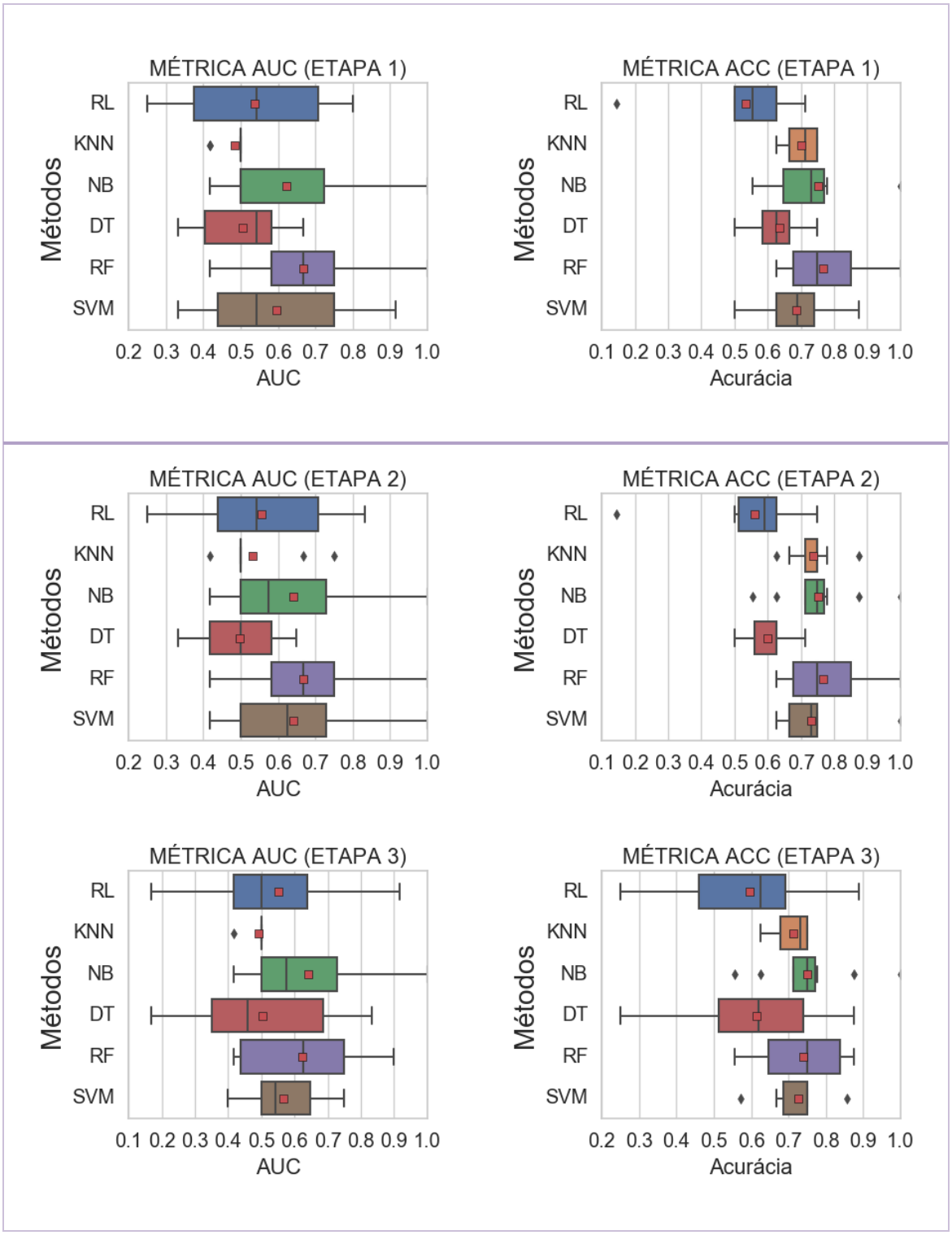


Figura 45: Métricas de AUC e acurácia (ACC) para série de treinamento do poço W4.

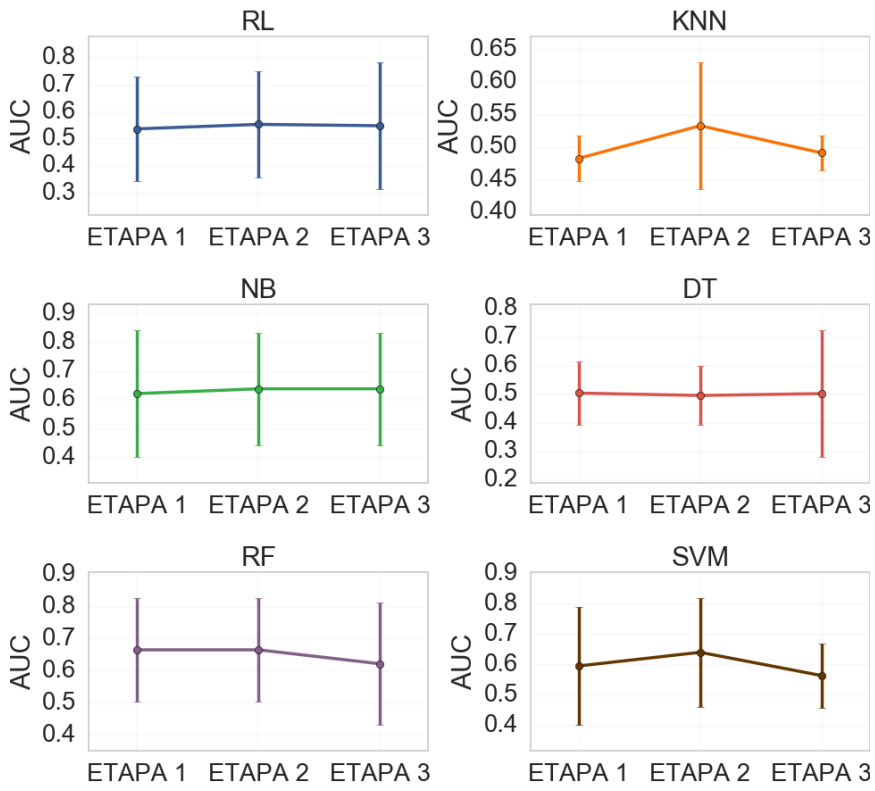


Figura 46: Médias de AUC com desvio-padrão obtidas em cada etapa da série de treinamento para poço W4.

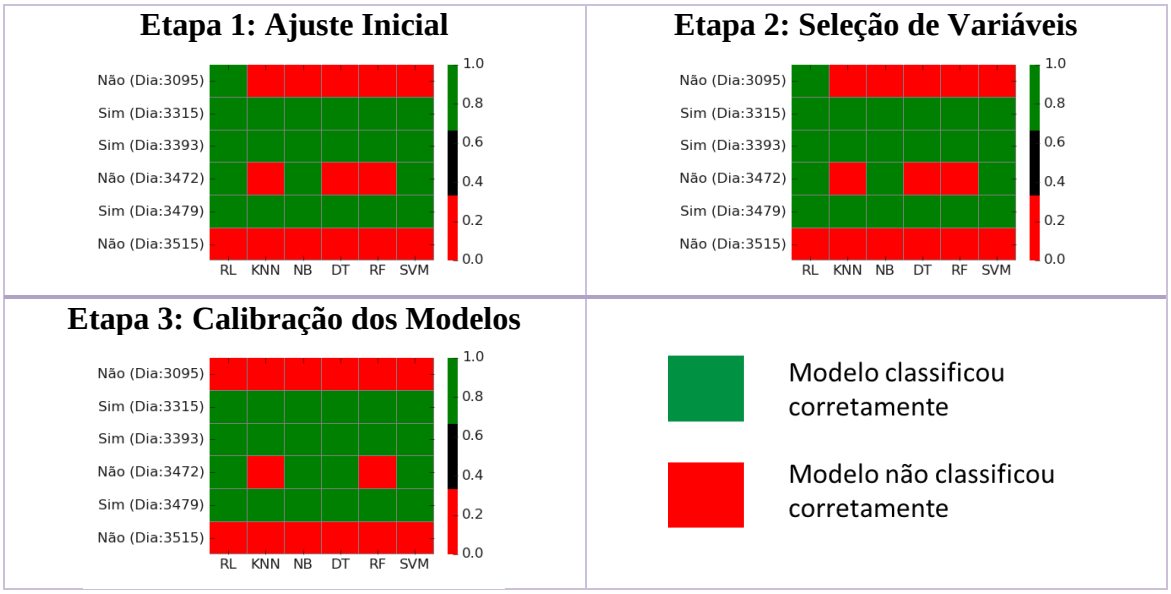


Figura 47: Resultados obtidos na série de validação para cada um dos métodos nas três etapas analisadas. Poço W4.

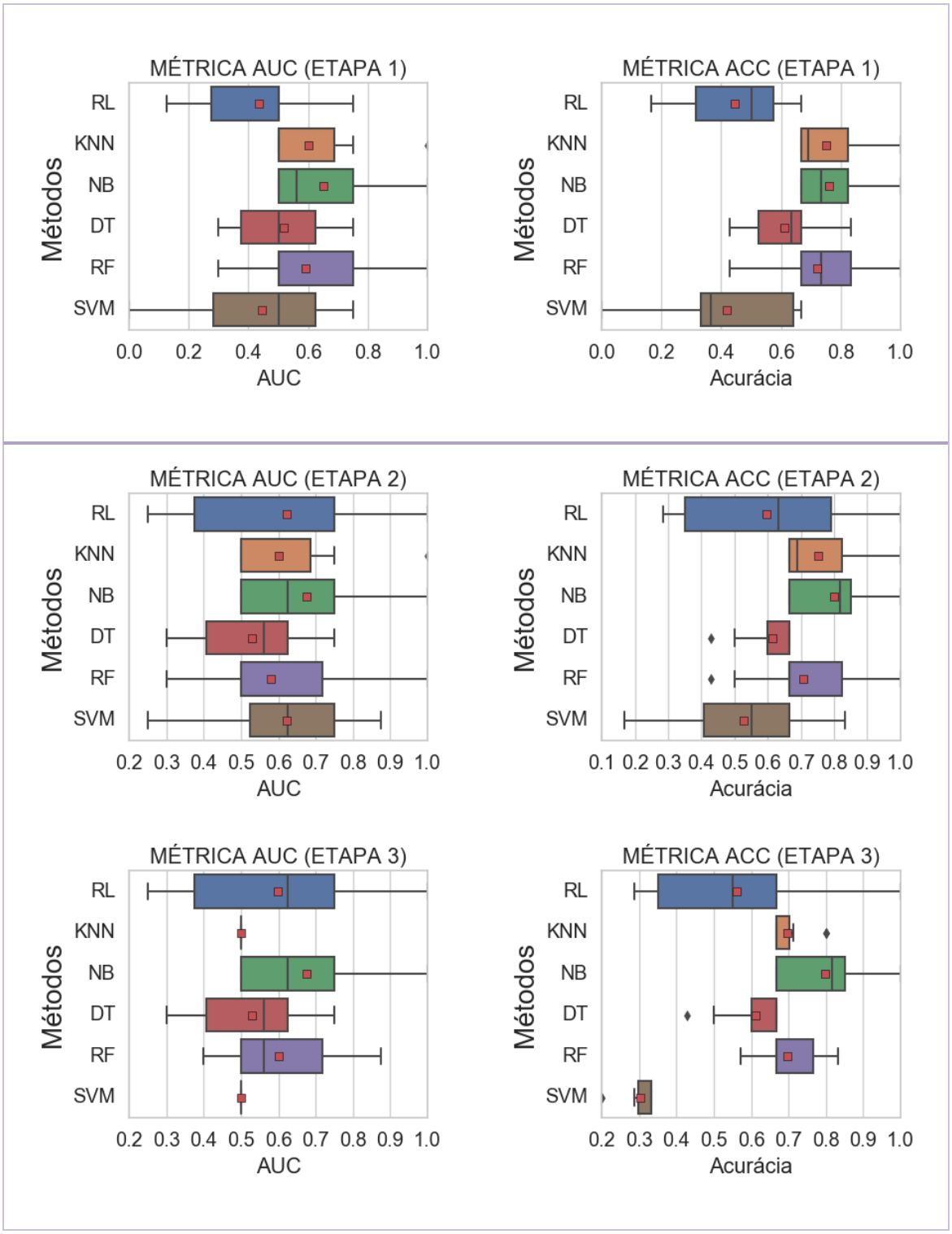


Figura 48: Métricas de AUC e acurácia (ACC) para série de treinamento do poço W5.

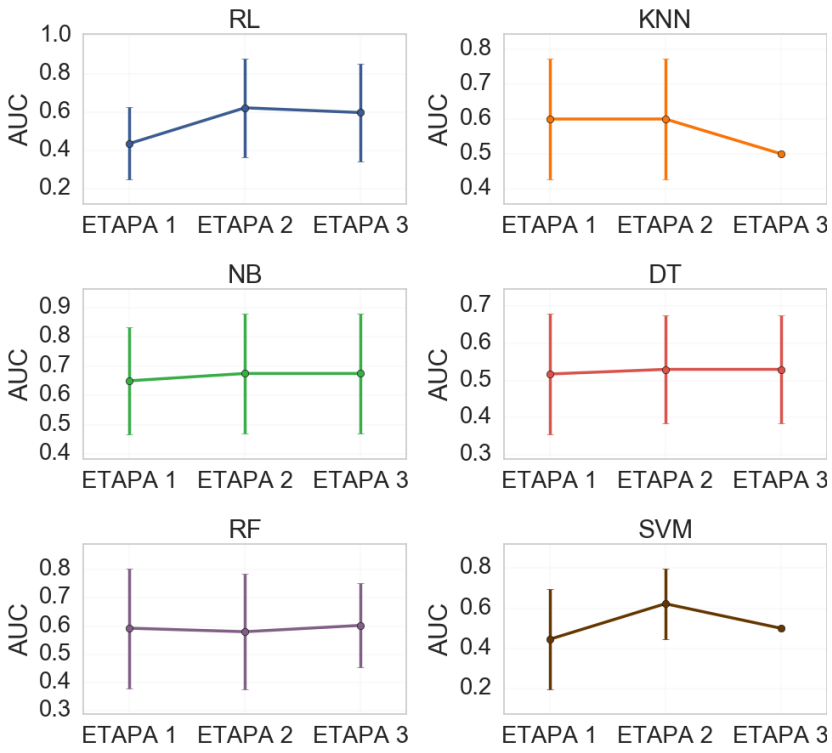


Figura 49: Médias de AUC com desvio-padrão obtidas em cada etapa da série de treinamento para poço W5.

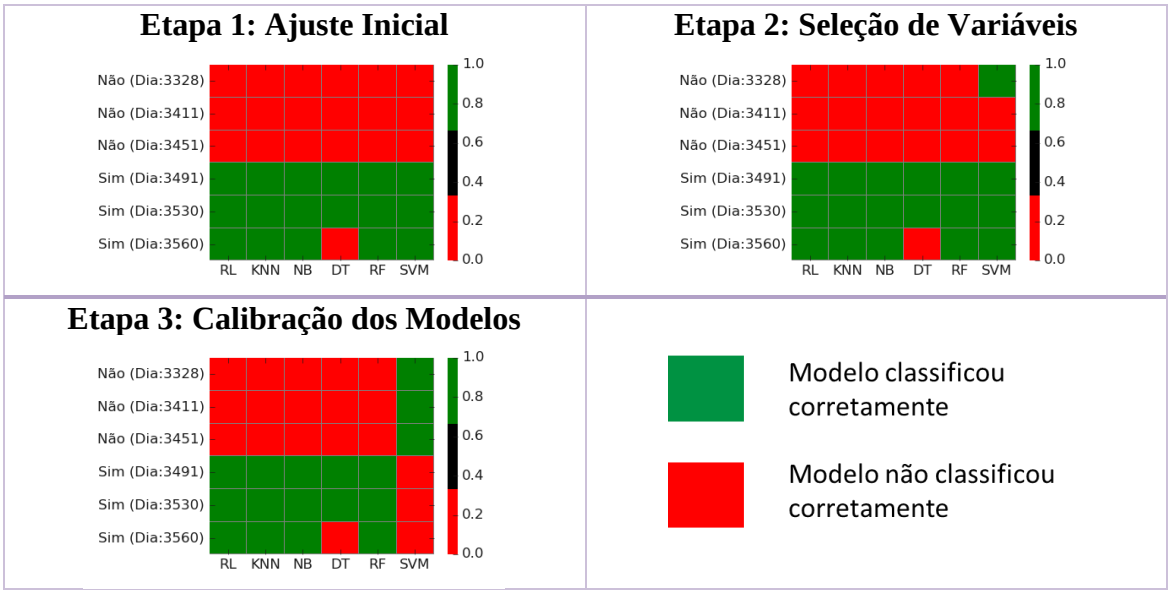


Figura 50: Resultados obtidos na série de validação para cada um dos métodos nas três etapas analisadas. Poço W5.

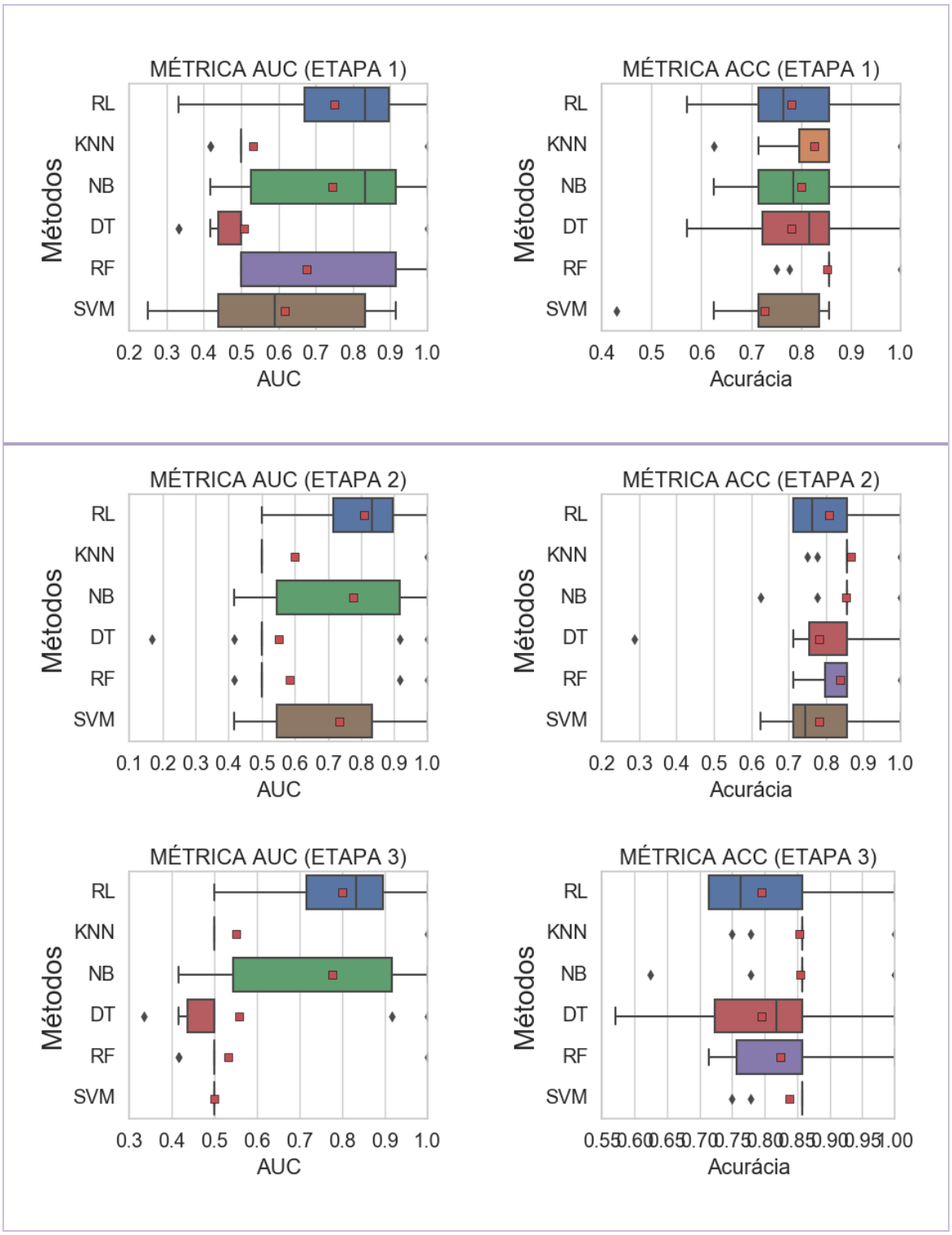


Figura 51: Métricas de AUC e acurácia (ACC) para série de treinamento do poço W6.

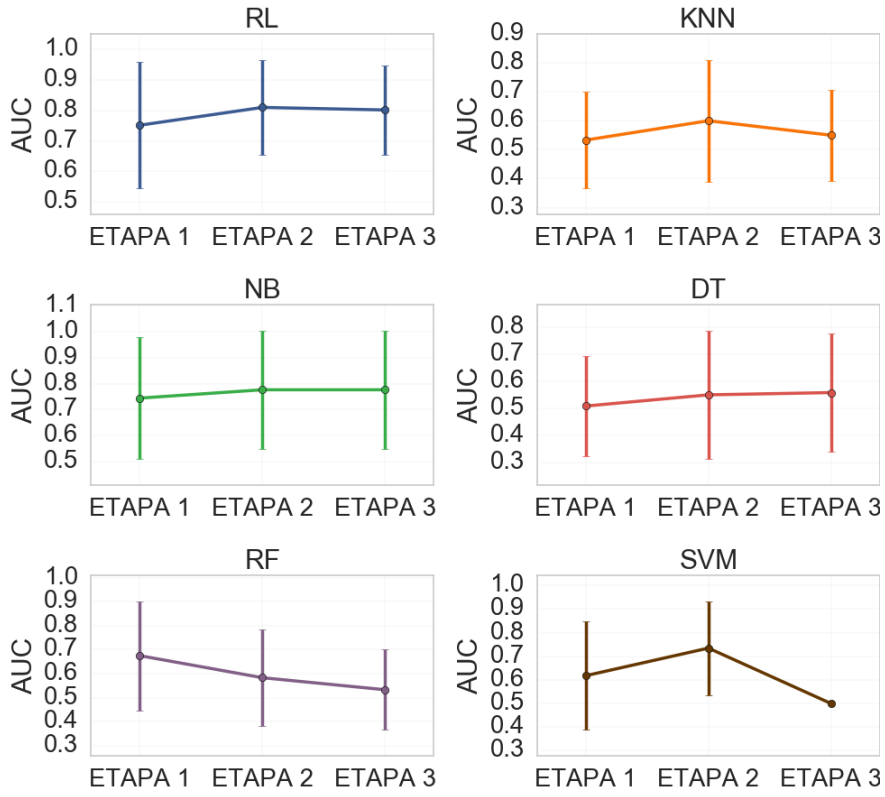


Figura 52: Médias de AUC com desvio-padrão obtidas em cada etapa da série de treinamento para poço W6.

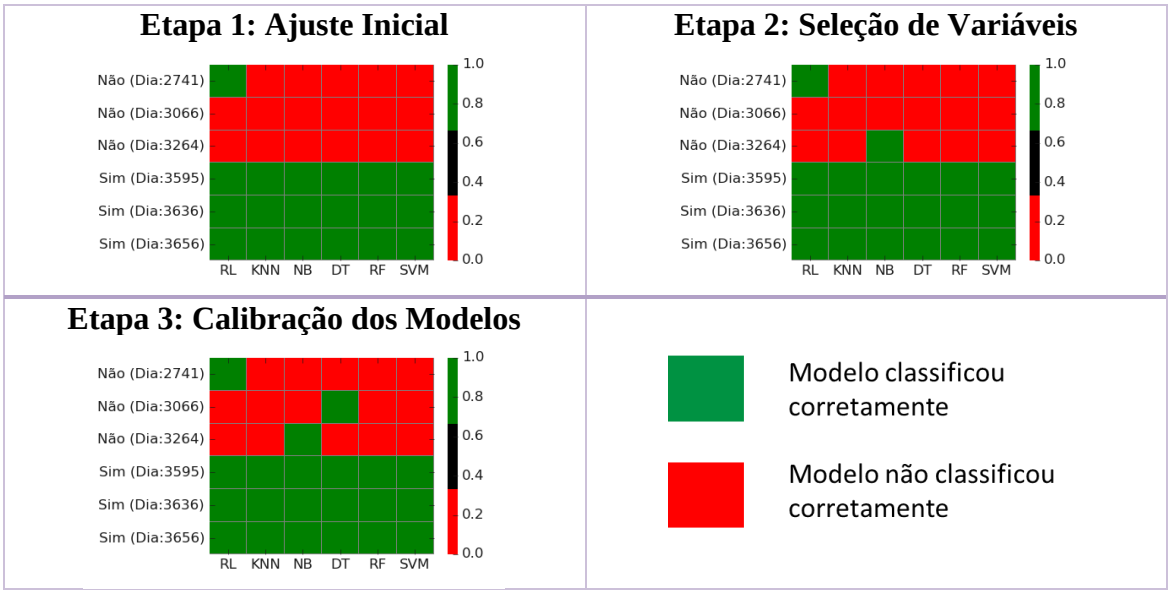


Figura 53: Resultados obtidos na série de validação para cada um dos métodos nas três etapas analisadas. Poço W6.

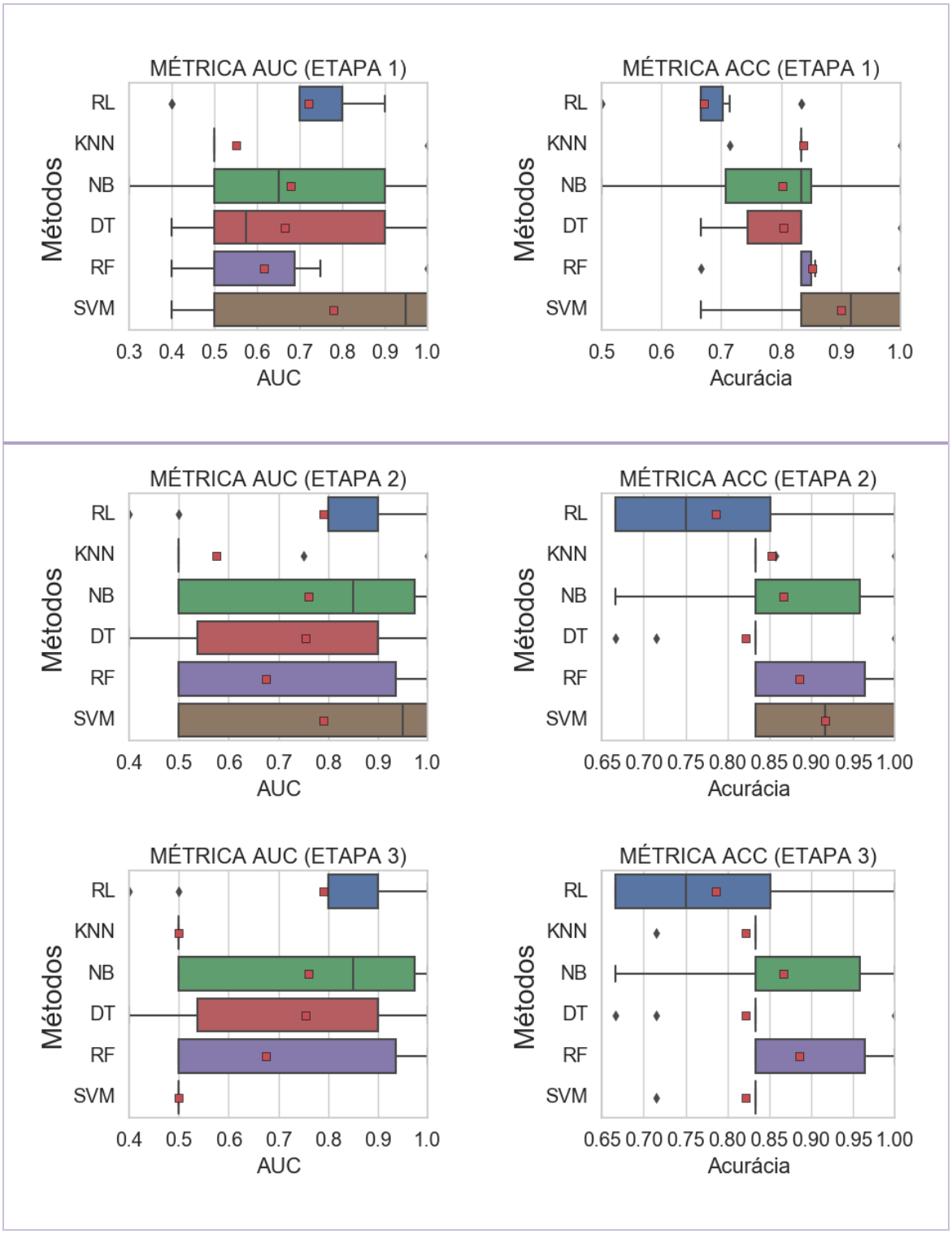


Figura 54: Métricas de AUC e acurácia (ACC) para série de treinamento do poço W7.

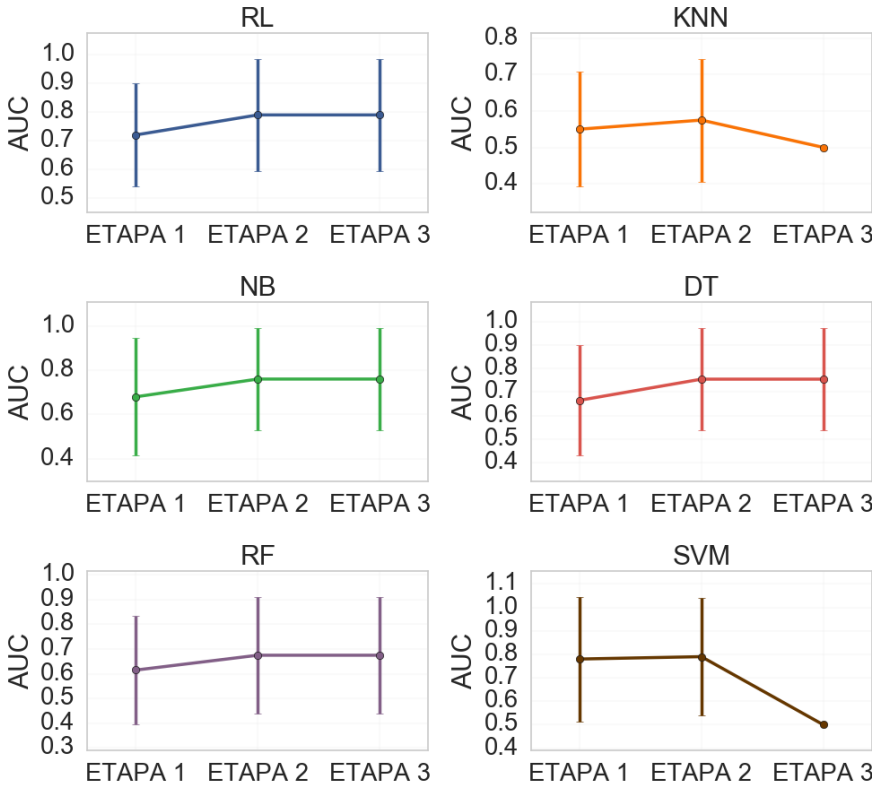


Figura 55: Médias de AUC com desvio-padrão obtidas em cada etapa da série de treinamento para poço W7.

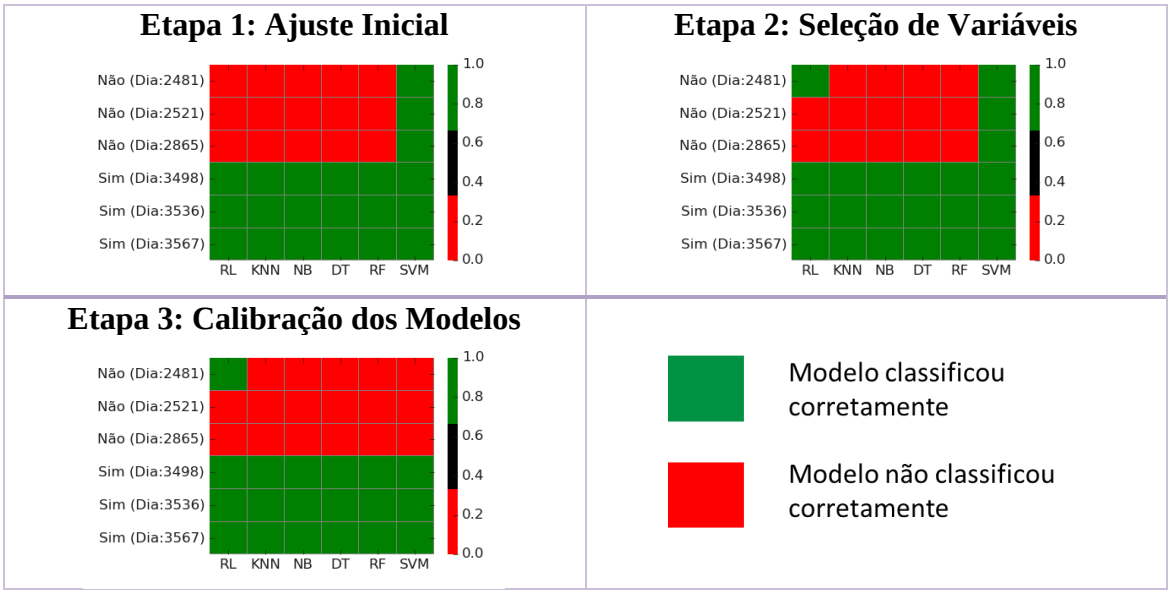


Figura 56: Resultados obtidos na série de validação para cada um dos métodos nas três etapas analisadas. Poço W7.

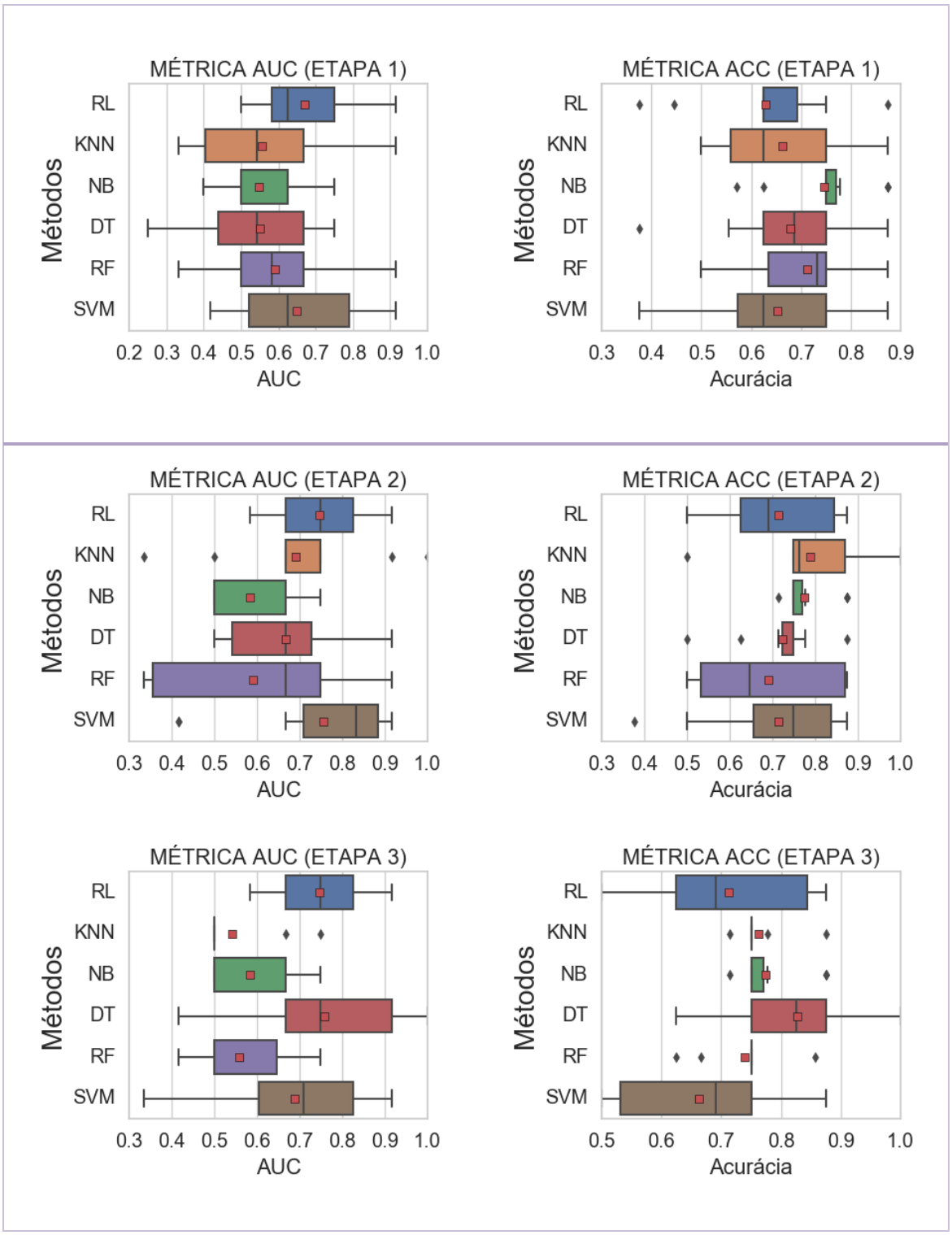


Figura 57: Métricas de AUC e acurácia (ACC) para série de treinamento do poço W8.

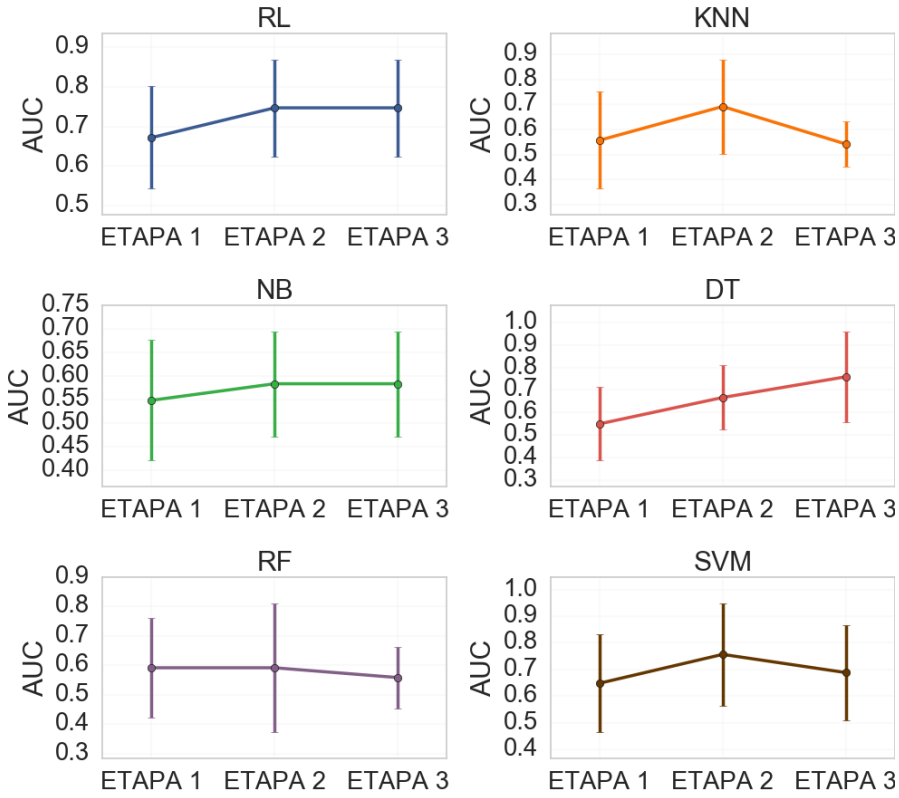


Figura 58: Médias de AUC com desvio-padrão obtidas em cada etapa da série de treinamento para poço W8.

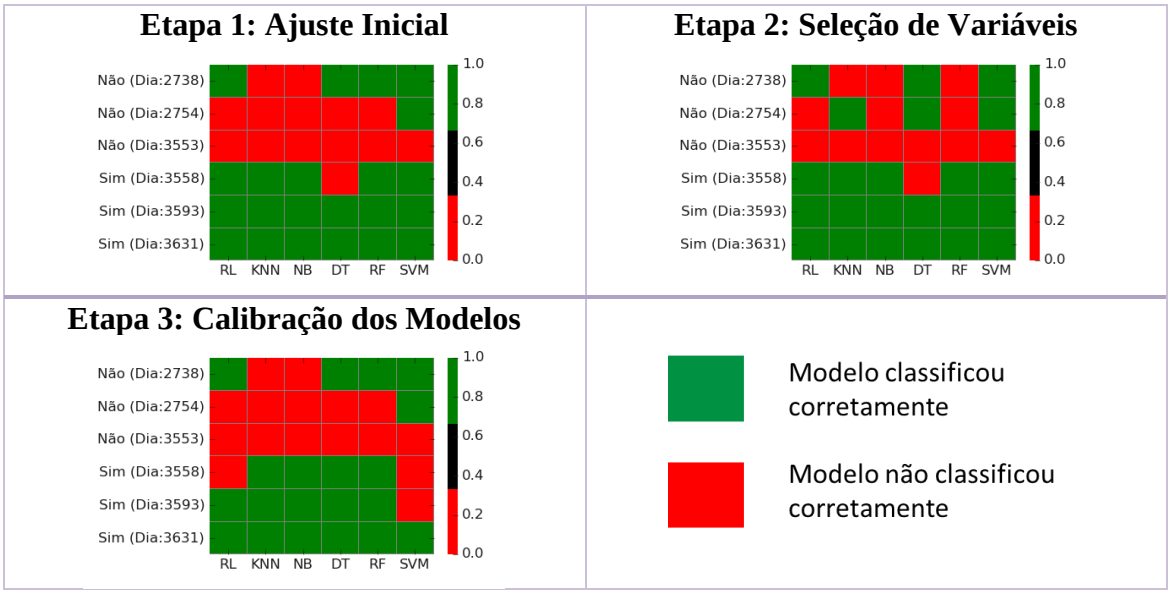


Figura 59: Resultados obtidos na série de validação para cada um dos métodos nas três etapas analisadas. Poço W8.

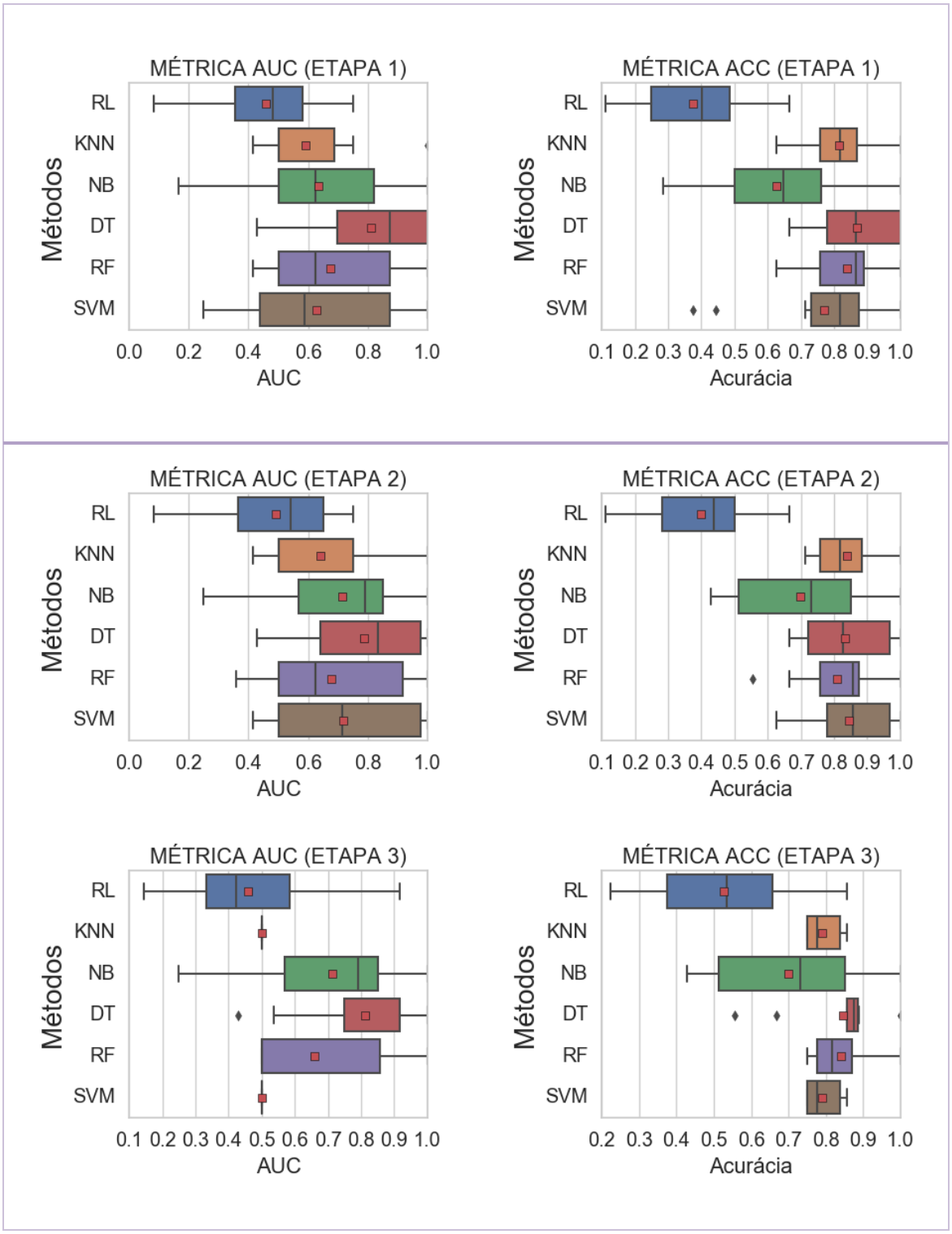


Figura 60: Métricas de AUC e acurácia (ACC) para série de treinamento do poço W9.

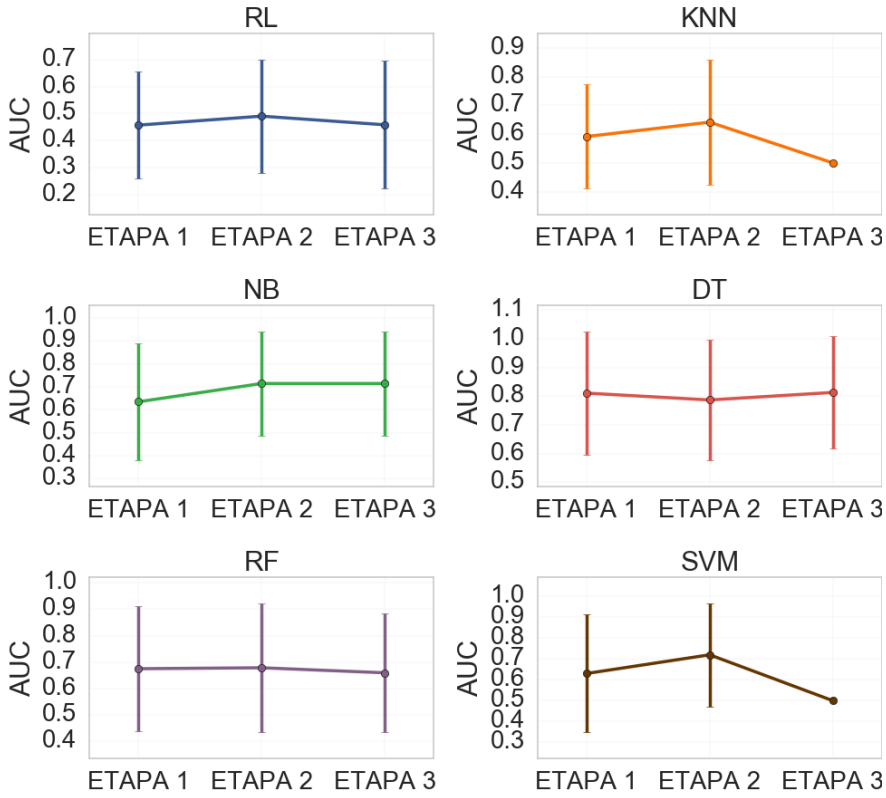


Figura 61: Médias de AUC com desvio-padrão obtidas em cada etapa da série de treinamento para poço W9.

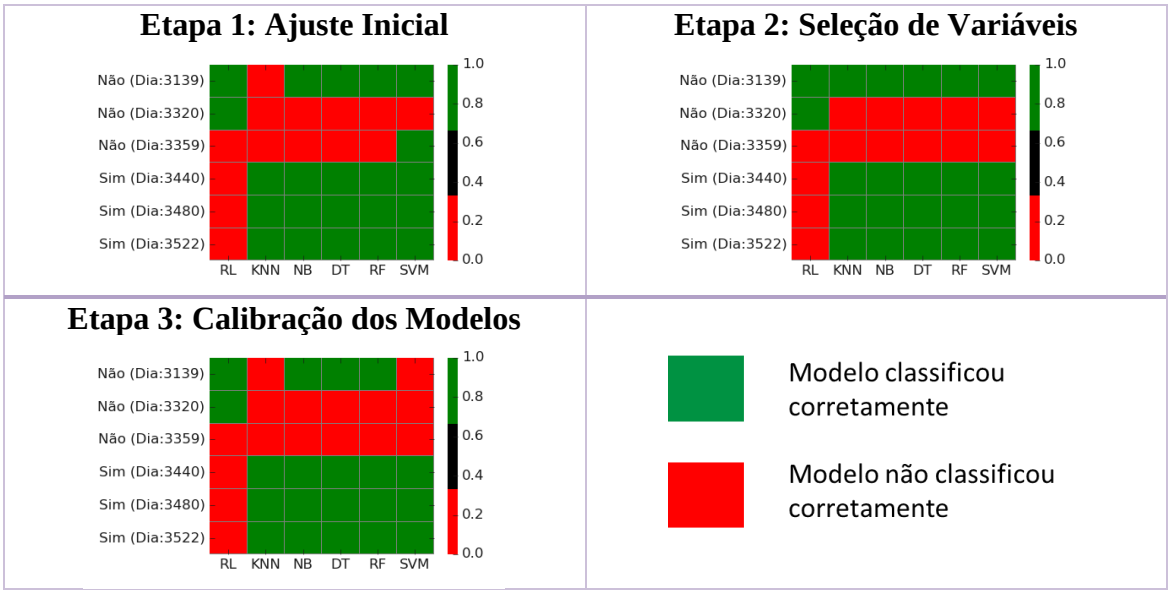


Figura 62: Resultados obtidos na série de validação para cada um dos métodos nas três etapas analisadas. Poço W9.

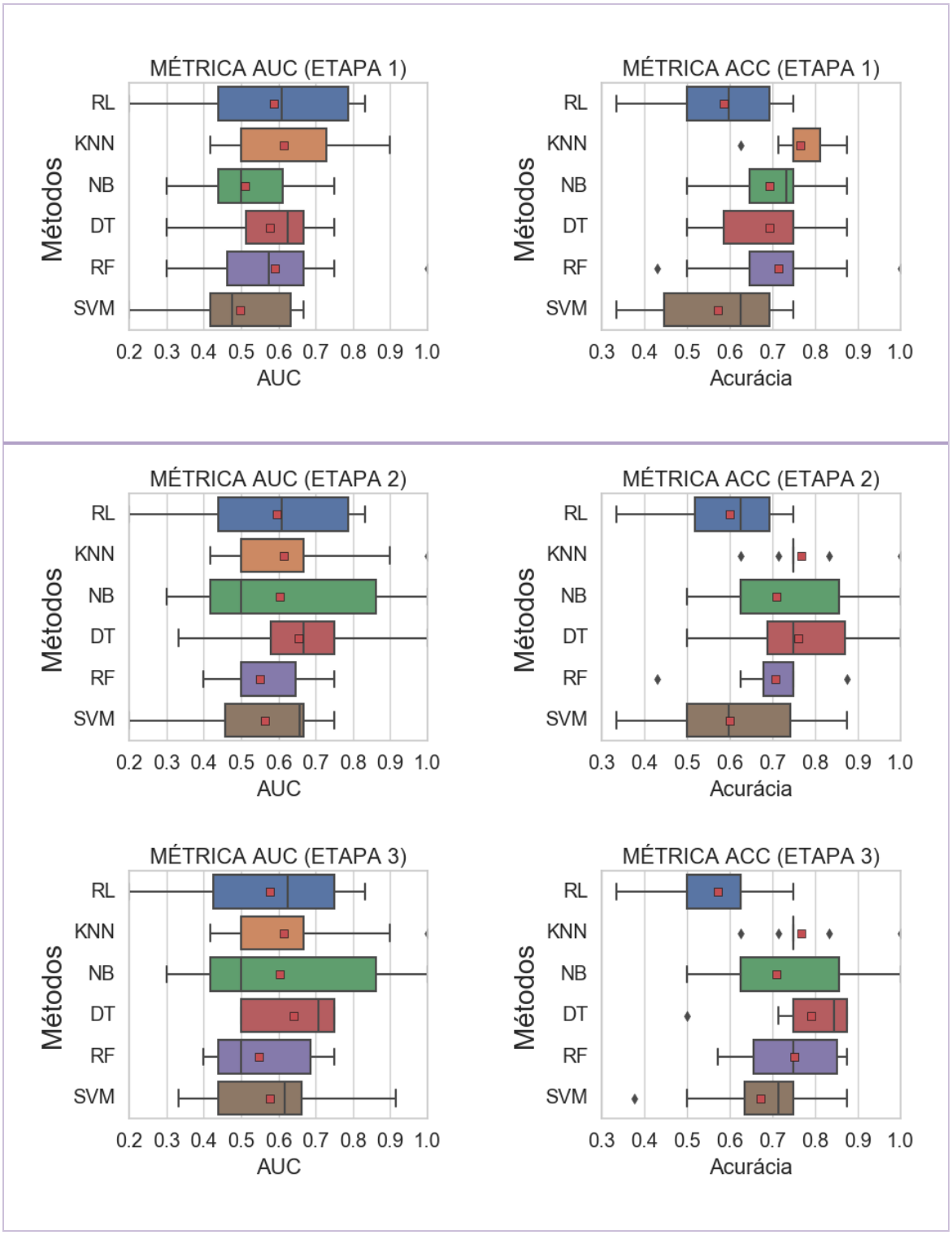


Figura 63: Métricas de AUC e acurácia (ACC) para série de treinamento do poço W10.

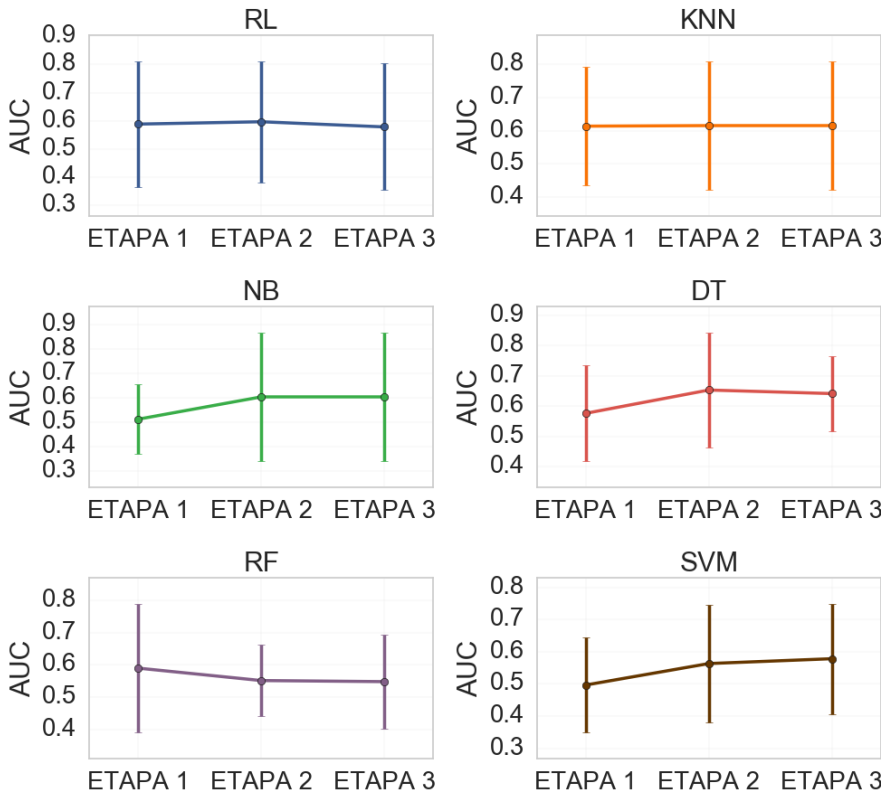


Figura 64: Médias de AUC com desvio-padrão obtidas em cada etapa da série de treinamento para poço W10.

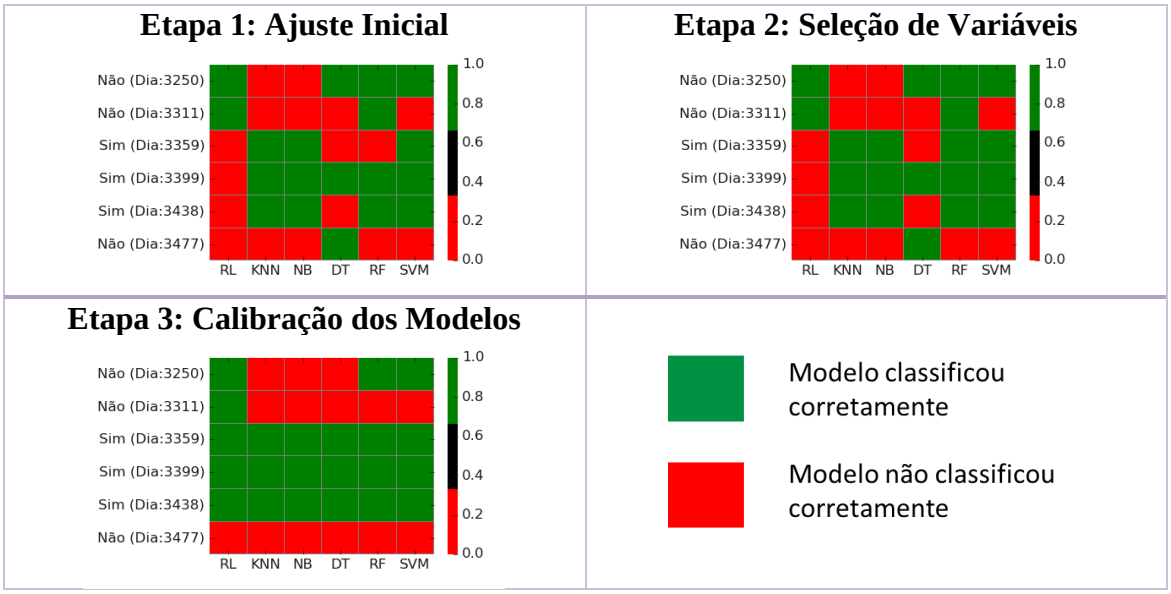


Figura 65: Resultados obtidos na série de validação para cada um dos métodos nas três etapas analisadas. Poço W10.

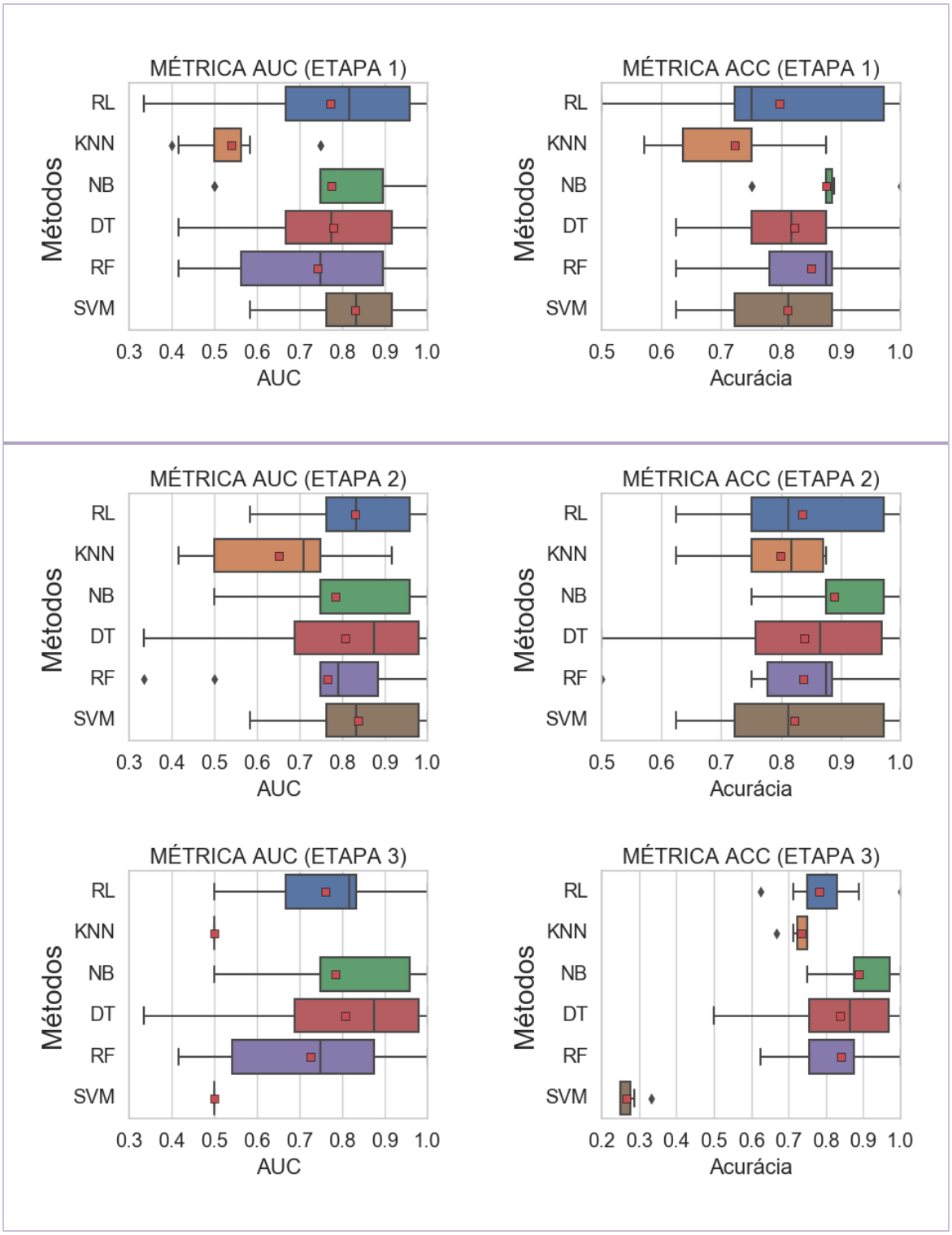


Figura 66: Métricas de AUC e acurácia (ACC) para série de treinamento do poço W11.

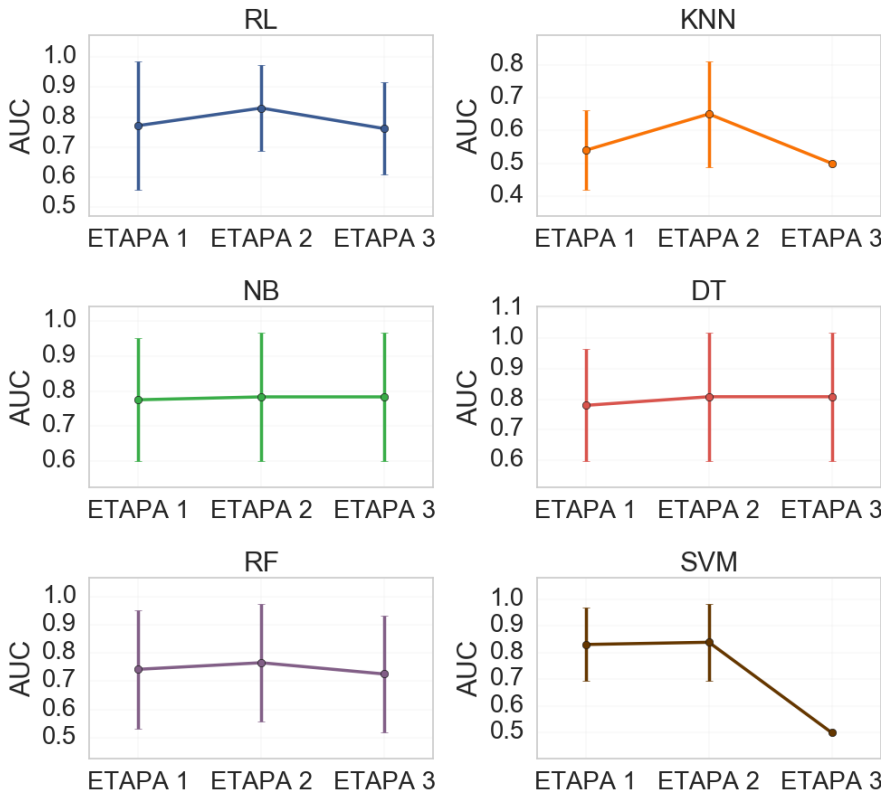


Figura 67: Médias de AUC com desvio-padrão obtidas em cada etapa da série de treinamento para poço W11.

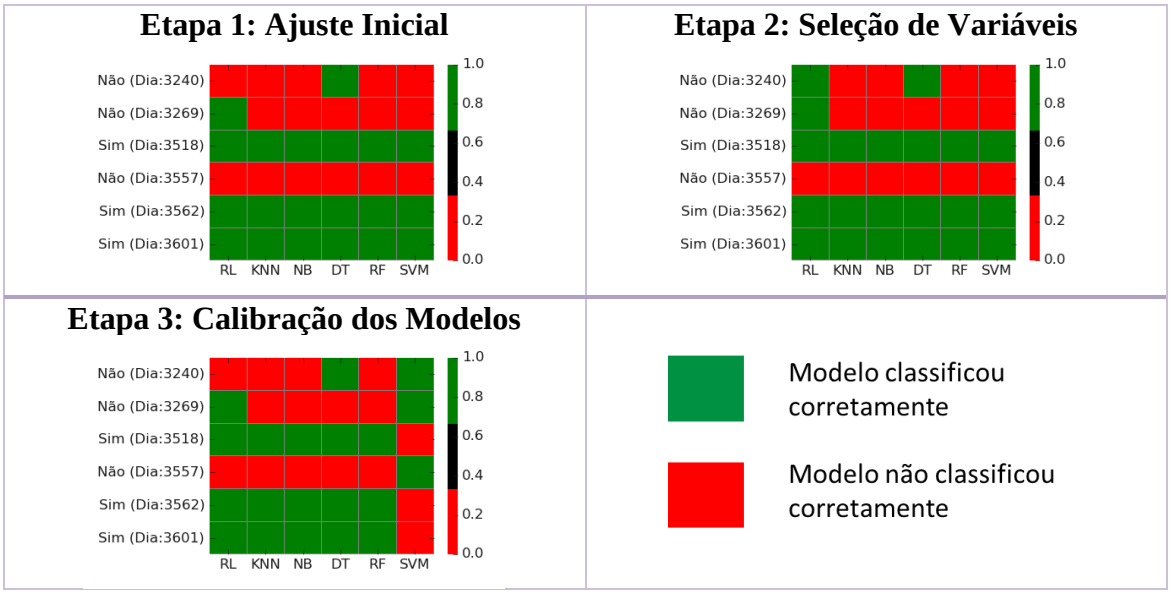


Figura 68: Resultados obtidos na série de validação para cada um dos métodos nas três etapas analisadas. Poço W11.

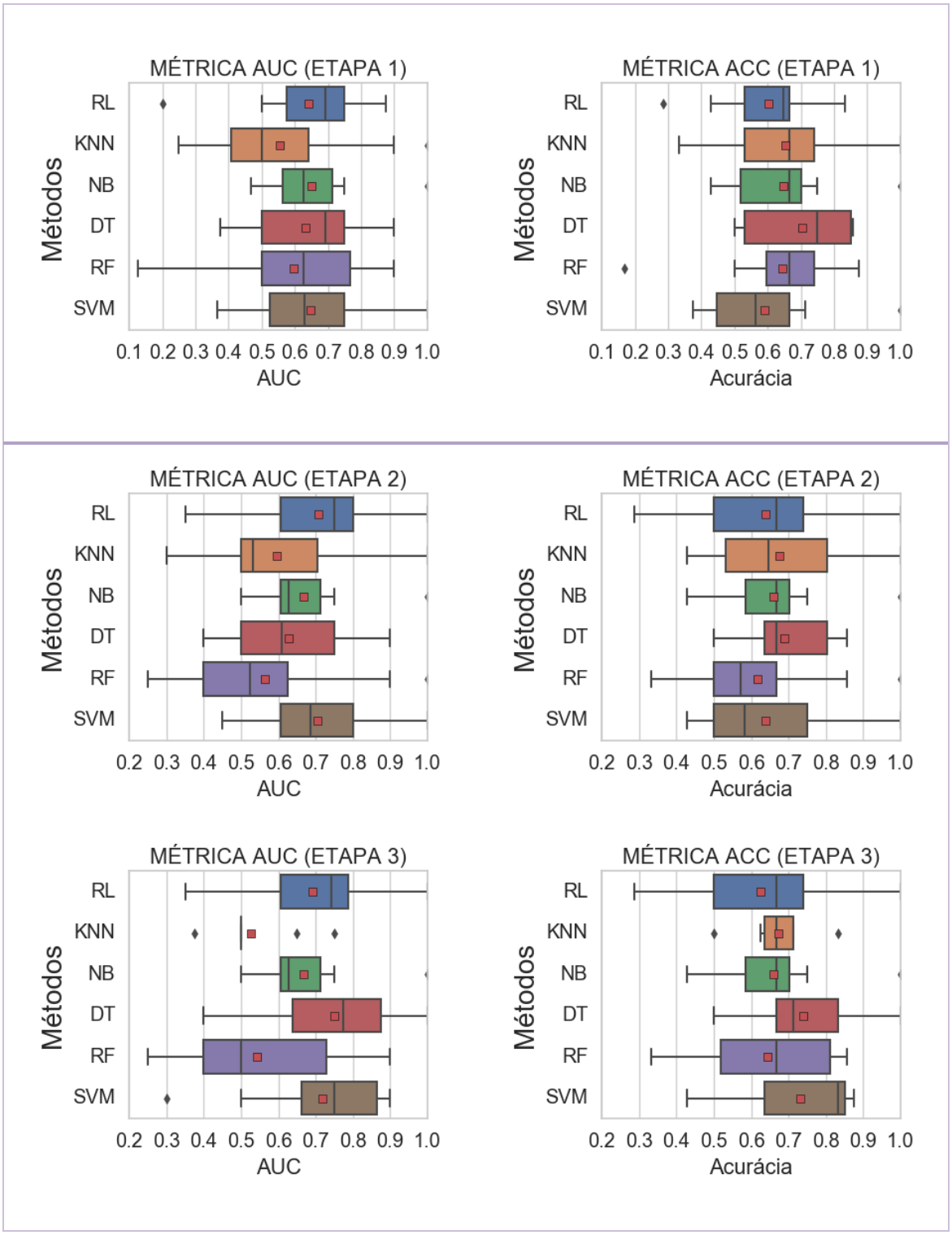


Figura 69: Métricas de AUC e acurácia (ACC) para série de treinamento do poço W12.

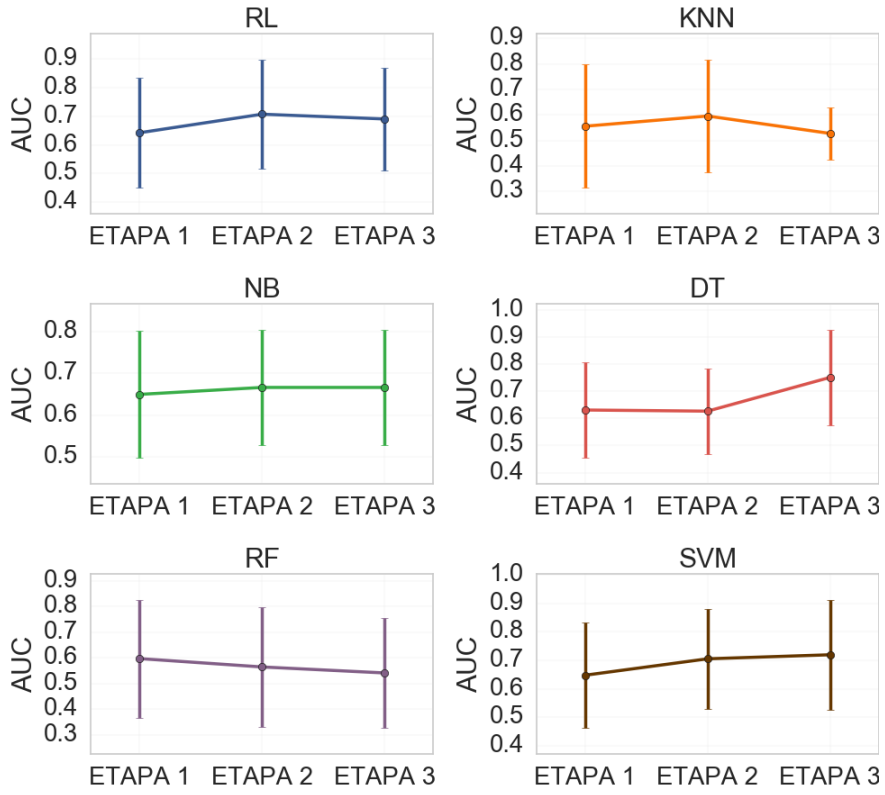


Figura 70: Médias de AUC com desvio-padrão obtidas em cada etapa da série de treinamento para poço W12.

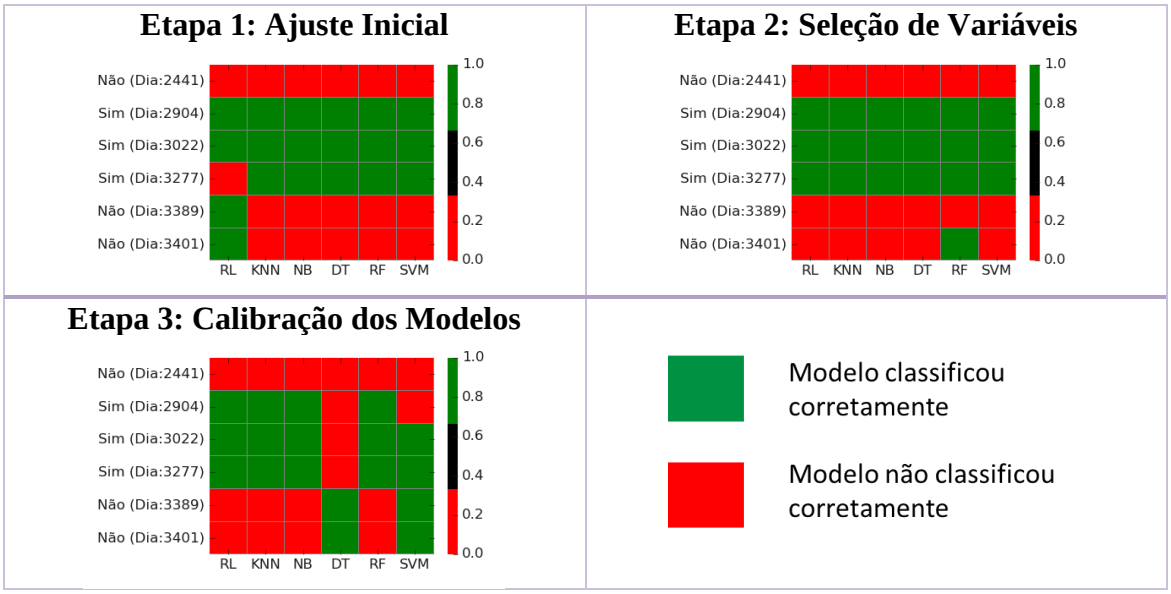


Figura 71: Resultados obtidos na série de validação para cada um dos métodos nas três etapas analisadas. Poço W12.

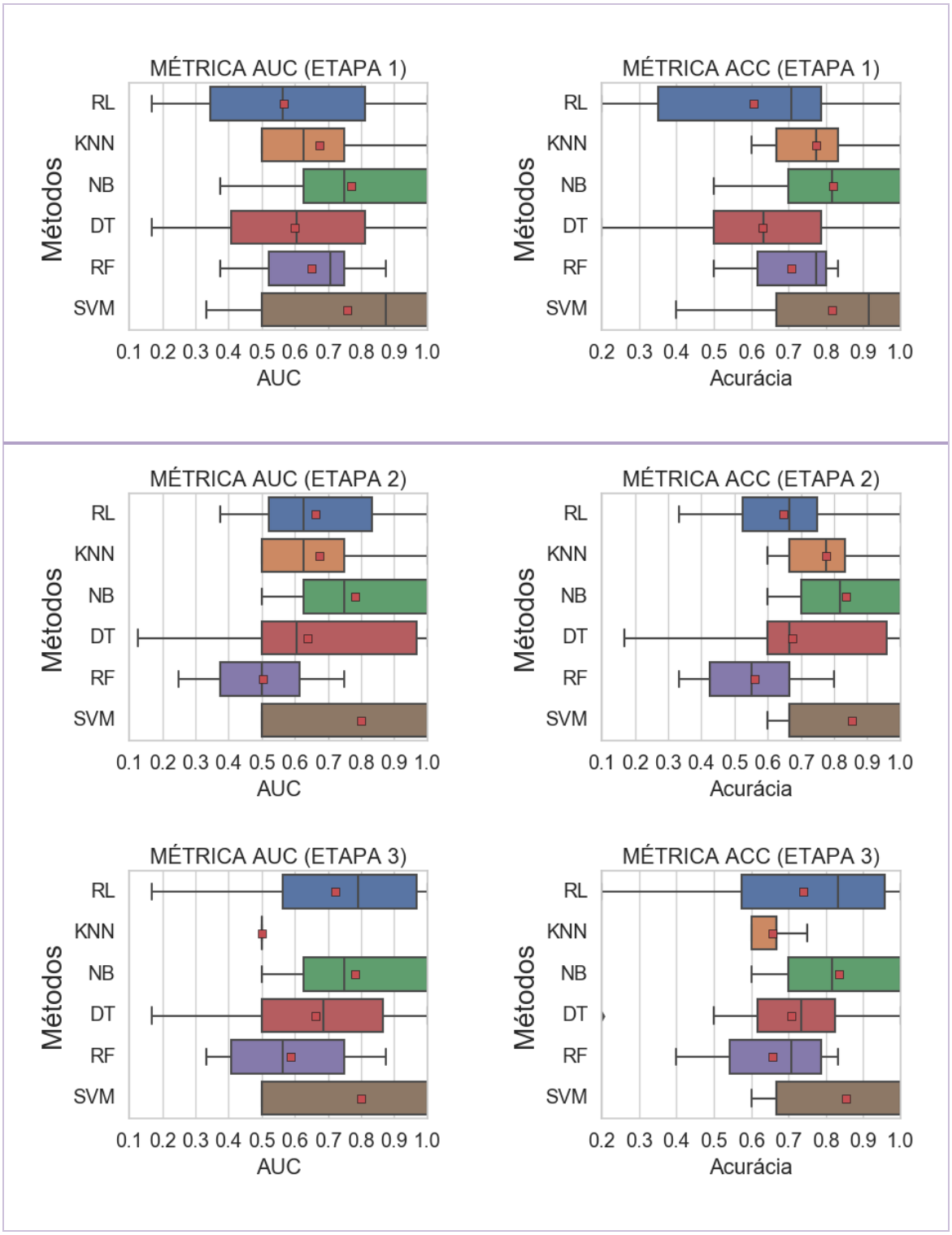


Figura 72: Métricas de AUC e acurácia (ACC) para série de treinamento do poço W13.

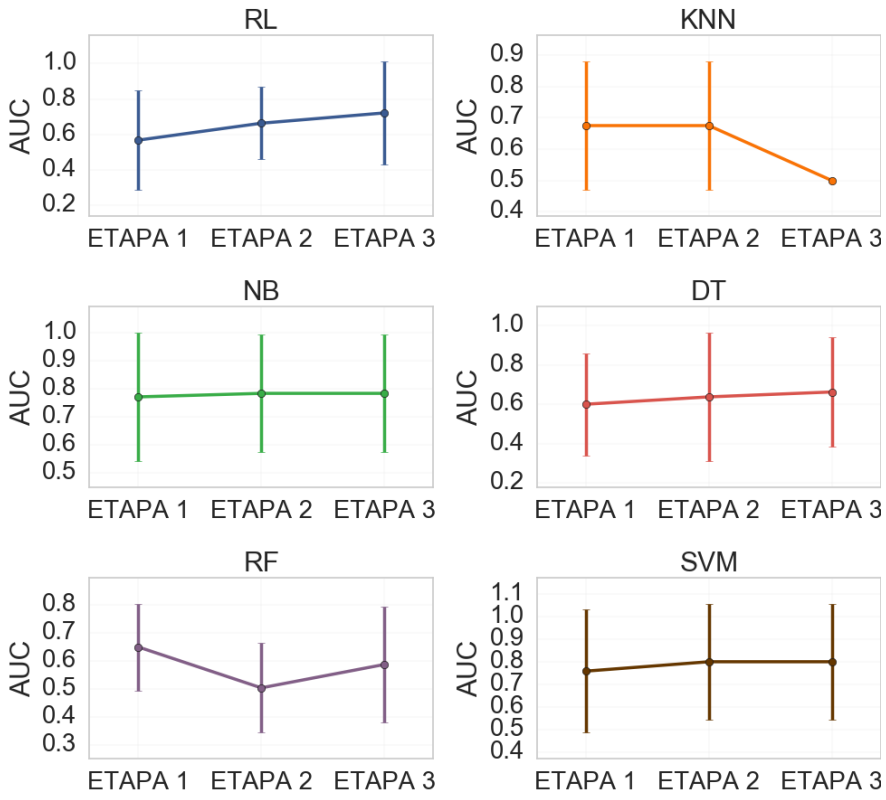


Figura 73: Médias de AUC com desvio-padrão obtidas em cada etapa da série de treinamento para poço W13.

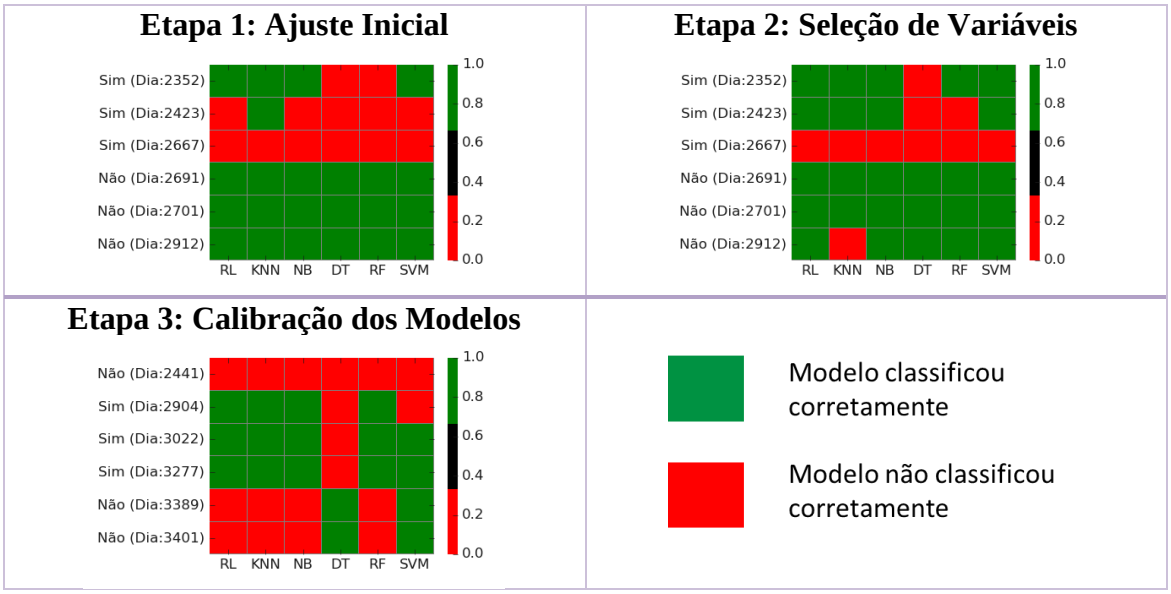
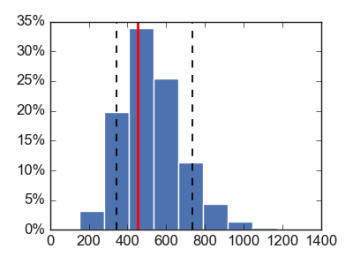


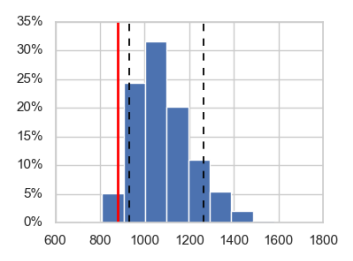
Figura 74: Resultados obtidos na série de validação para cada um dos métodos nas três etapas analisadas. Poço W13.

APÊNDICE II – RESULTADOS DA ETAPA DE REGRESSÃO

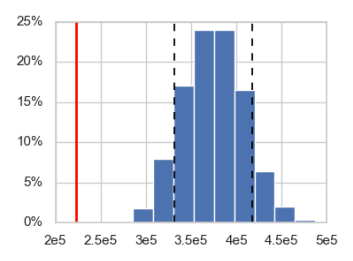
Os resultados a seguir mostram as vazões de óleo, água e gás, previstas para as 4 datas de cada poço. A linha em vermelho é o valor real obtido no teste, e as linhas pontilhadas, os limites, ou intervalos de previsão obtidos. Como os métodos RT e RFR obtiveram um desempenho ruim, estes não são apresentados a seguir.



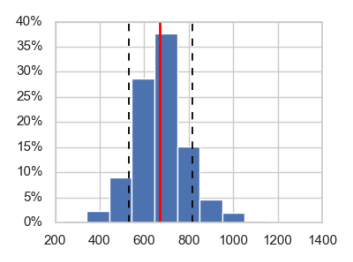
Vazão de Óleo – Data 1



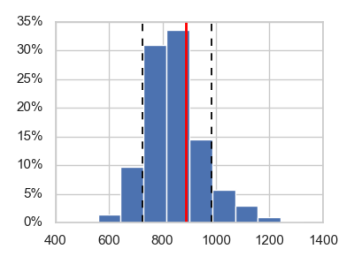
Vazão de Água – Data 1



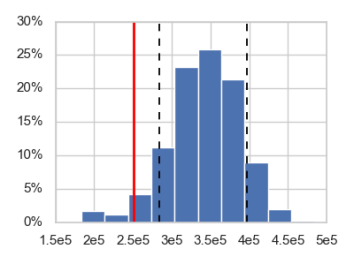
Vazão de Gás total – Data 1



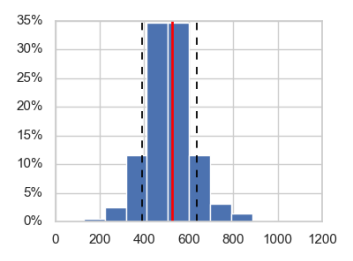
Vazão de Óleo – Data 2



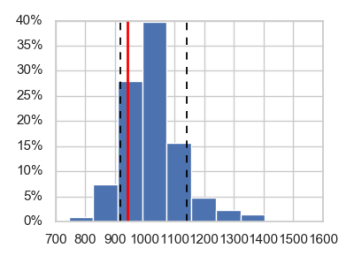
Vazão de Água – Data 2



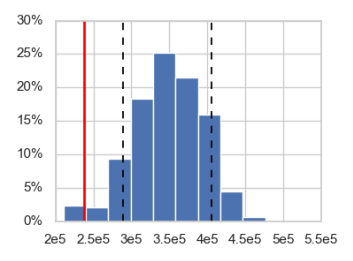
Vazão de Gás total – Data 2



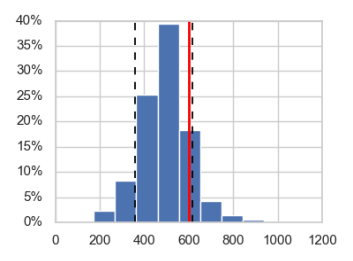
Vazão de Óleo – Data 3



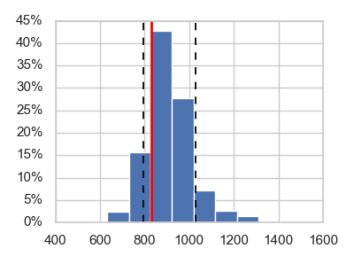
Vazão de Água – Data 3



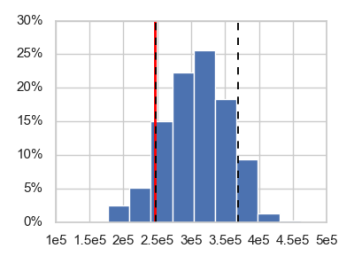
Vazão de Gás total – Data 3



Vazão de Óleo – Data 4



Vazão de Água – Data 4



Vazão de Gás total – Data 4

Figura 75: Resultados previstos para modelo MLR. Poço W1.

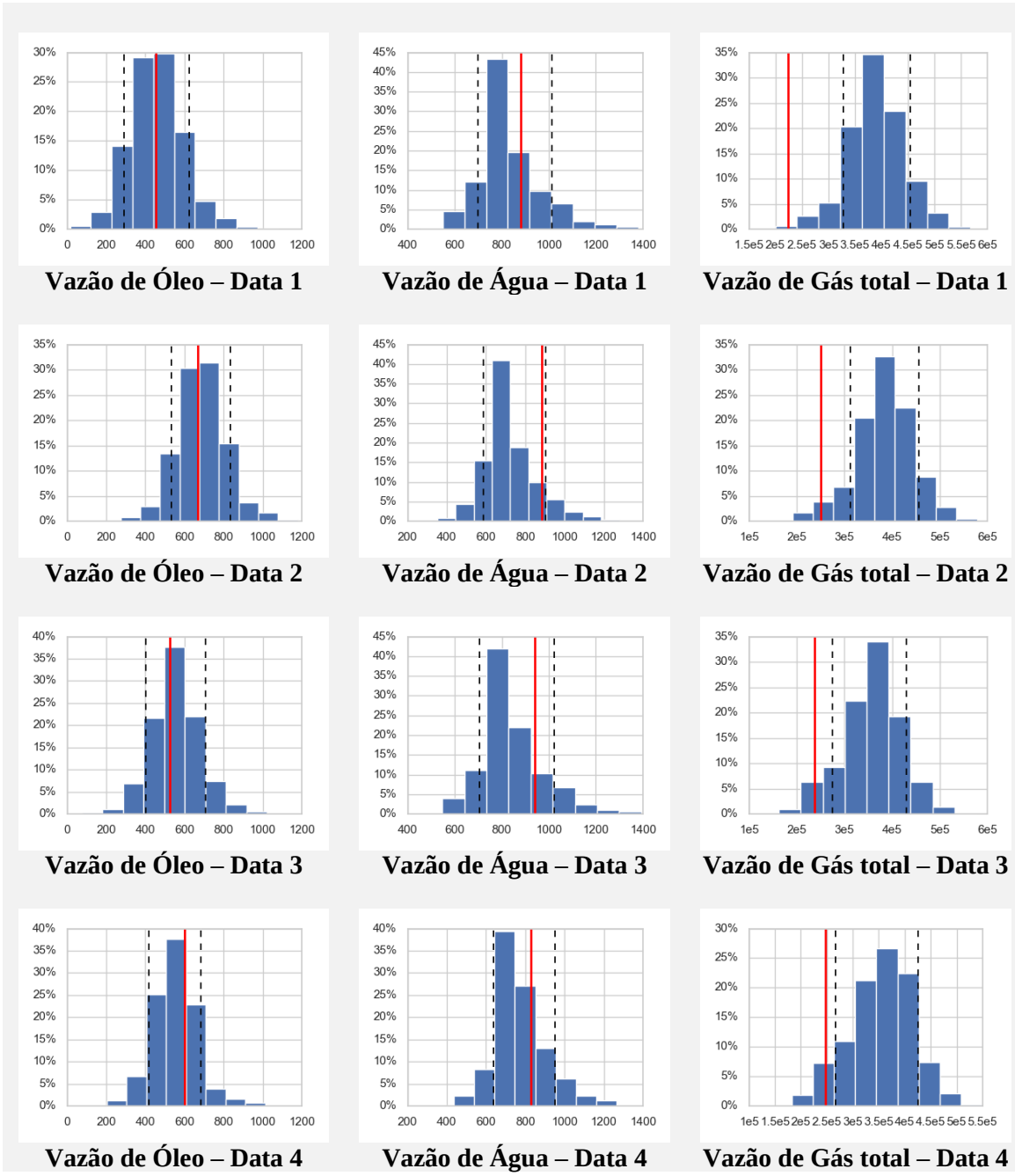


Figura 76: Resultados previstos para modelo SVR. Poço W1.

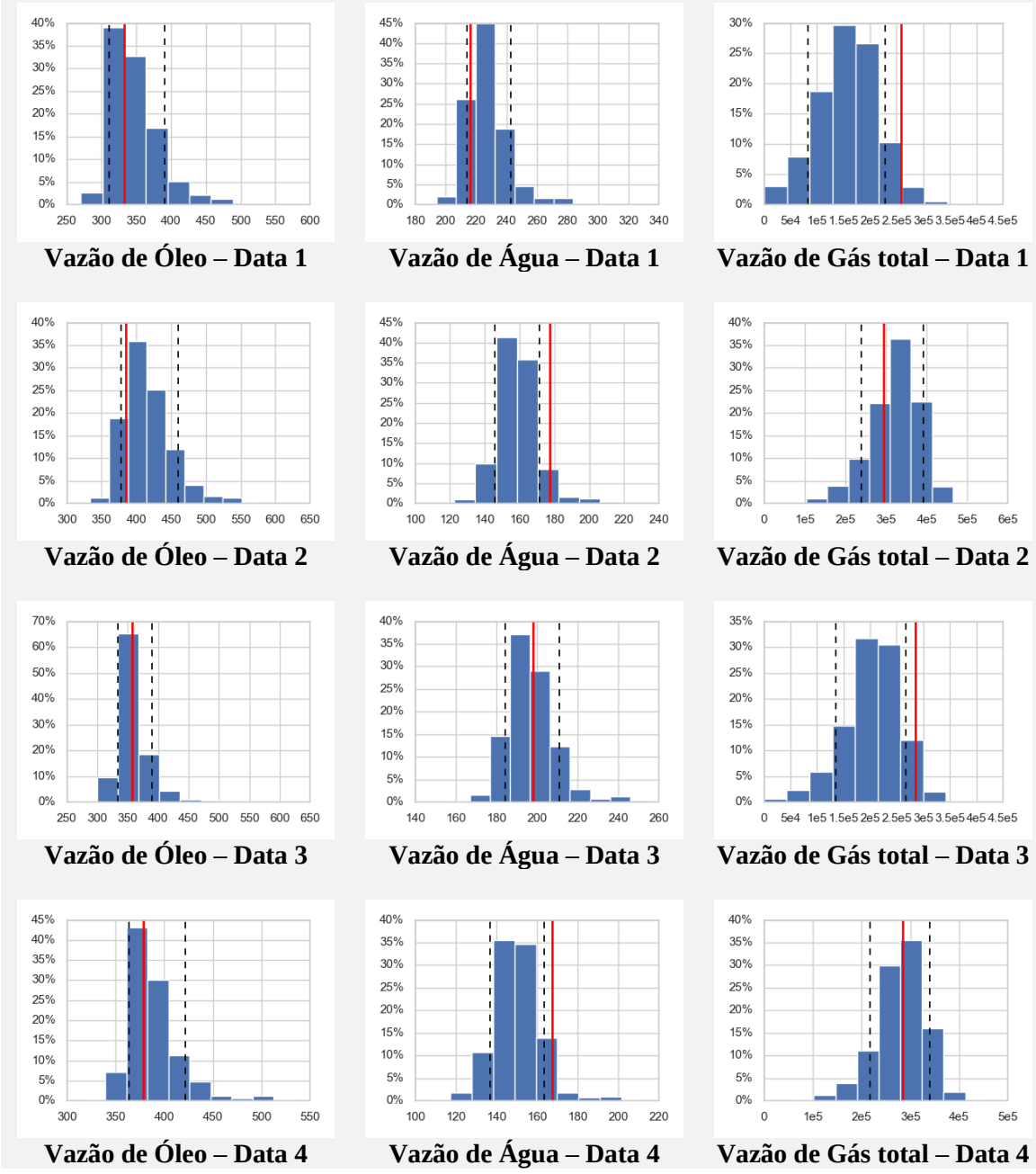
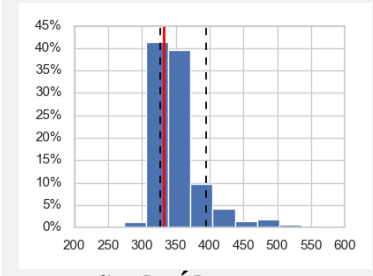
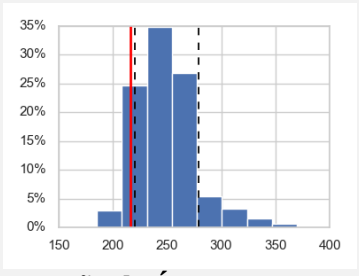


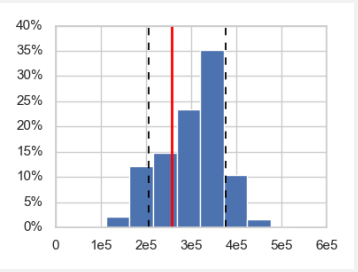
Figura 77: Resultados previstos para modelo MLR. Poço W2.



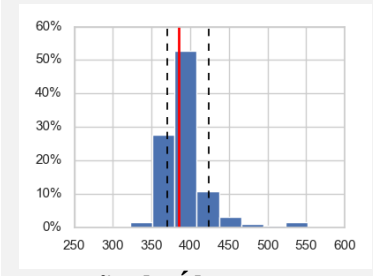
Vazão de Óleo – Data 1



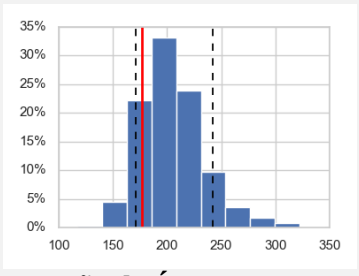
Vazão de Água – Data 1



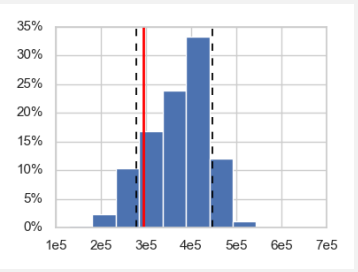
Vazão de Gás total – Data 1



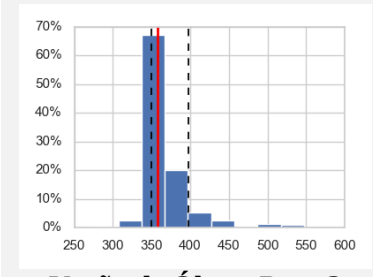
Vazão de Óleo – Data 2



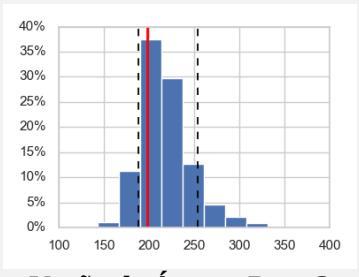
Vazão de Água – Data 2



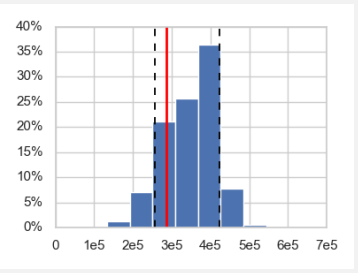
Vazão de Gás total – Data 2



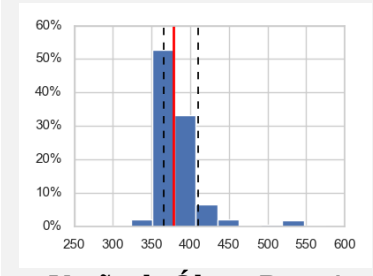
Vazão de Óleo – Data 3



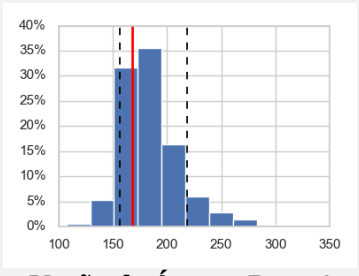
Vazão de Água – Data 3



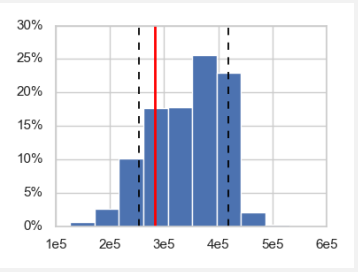
Vazão de Gás total – Data 3



Vazão de Óleo – Data 4



Vazão de Água – Data 4



Vazão de Gás total – Data 4

Figura 78: Resultados previstos para modelo SVR. Poço W2.

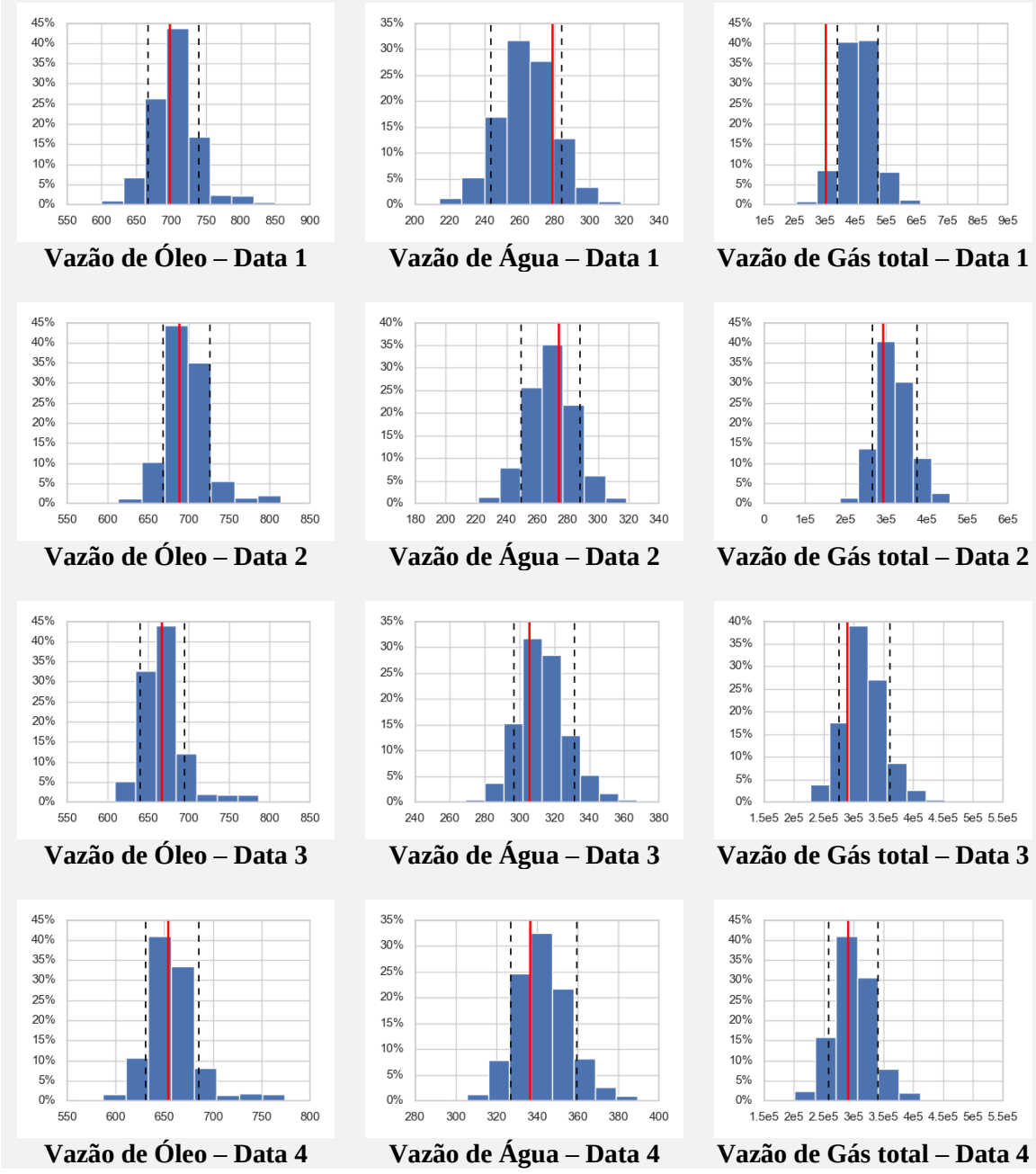


Figura 79: Resultados previstos para modelo MLR. Poço W3.

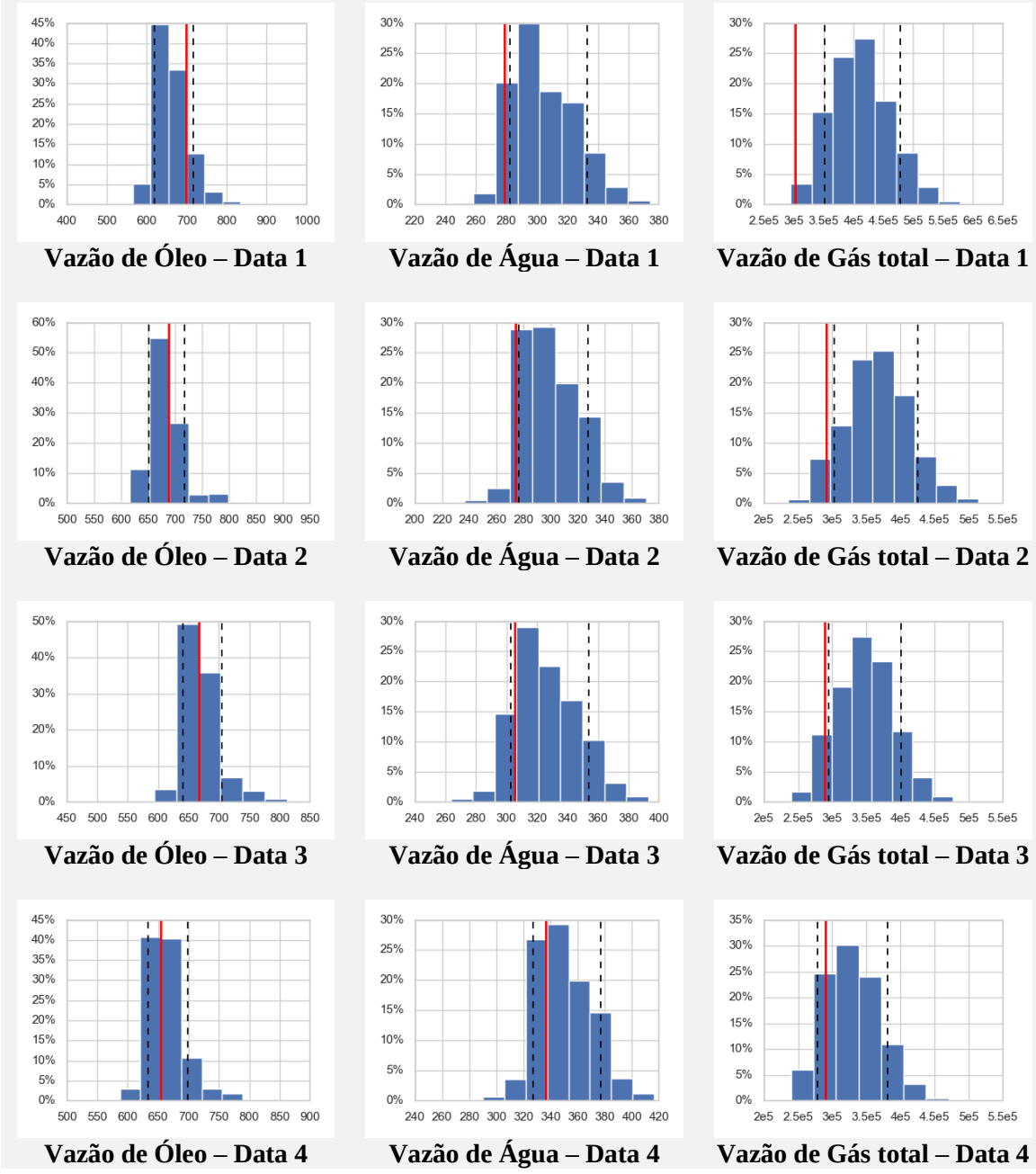


Figura 80: Resultados previstos para modelo SVR. Poço W3.

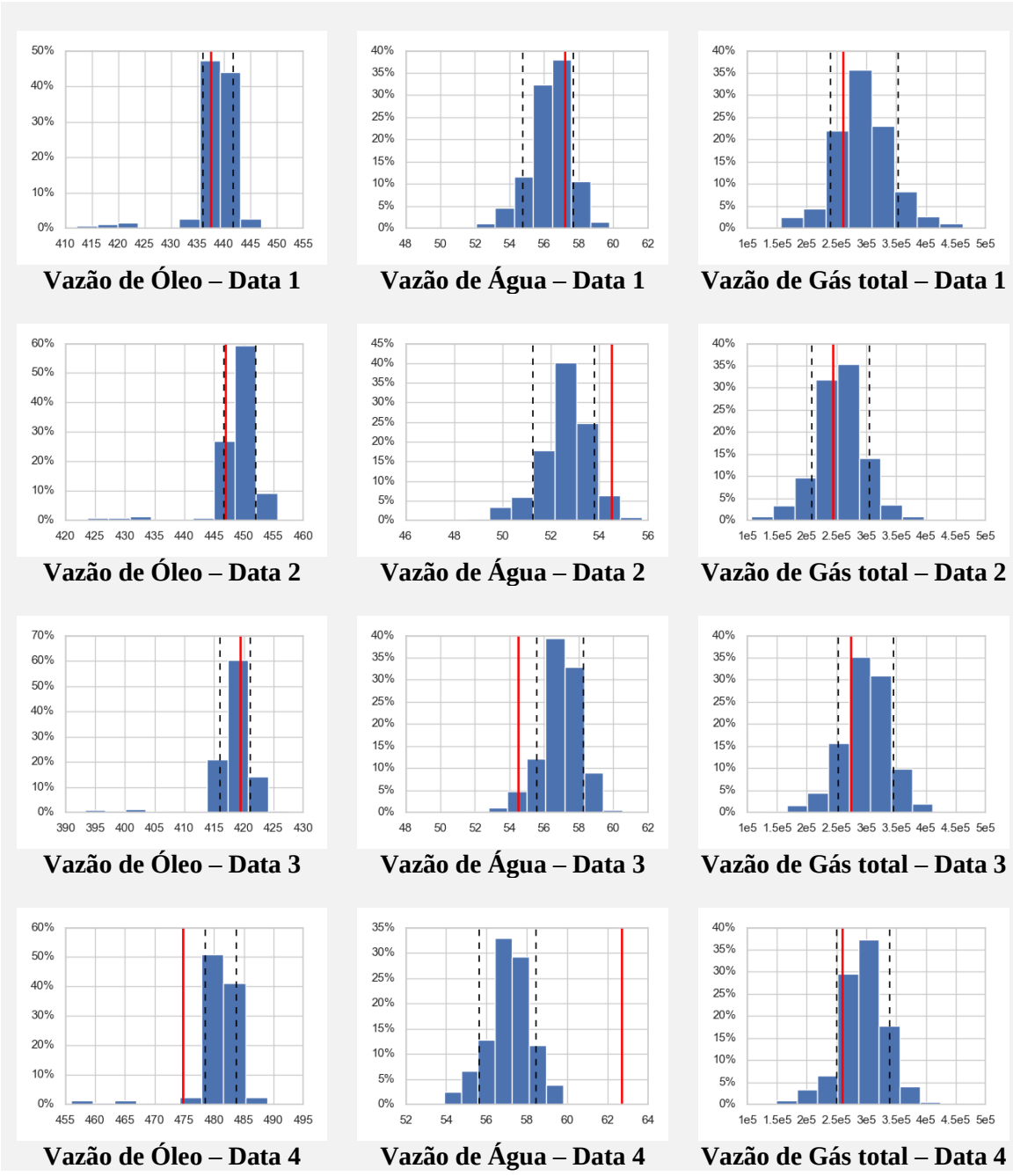
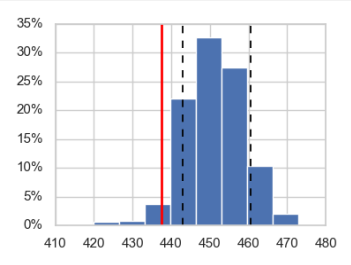
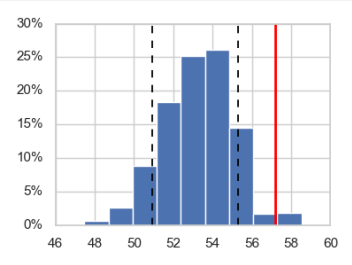


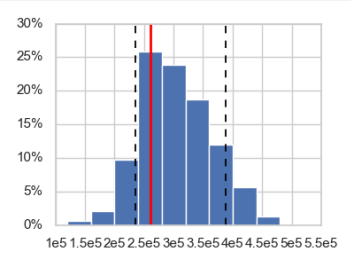
Figura 81: Resultados previstos para modelo MLR. Poço W4.



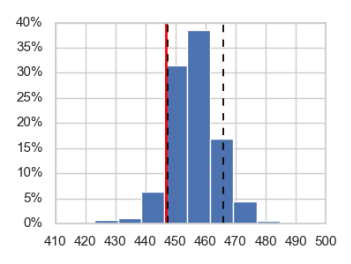
Vazão de Óleo – Data 1



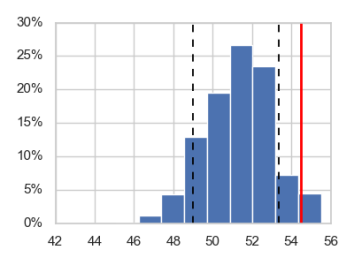
Vazão de Água – Data 1



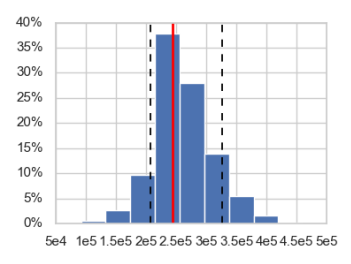
Vazão de Gás total – Data 1



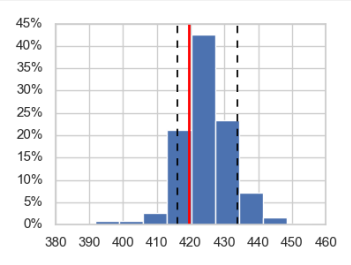
Vazão de Óleo – Data 2



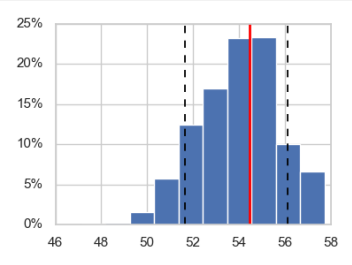
Vazão de Água – Data 2



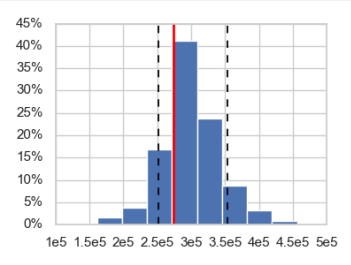
Vazão de Gás total – Data 2



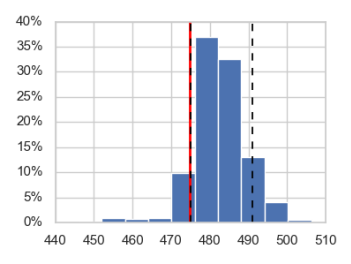
Vazão de Óleo – Data 3



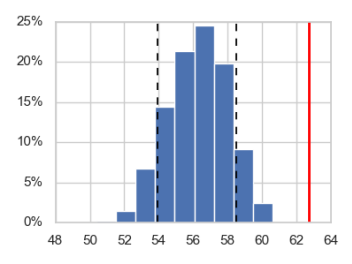
Vazão de Água – Data 3



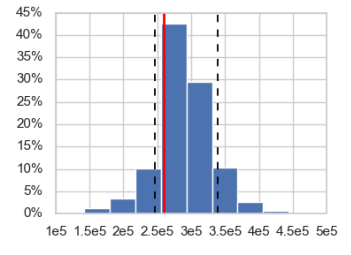
Vazão de Gás total – Data 3



Vazão de Óleo – Data 4



Vazão de Água – Data 4



Vazão de Gás total – Data 4

Figura 82: Resultados previstos para modelo SVR. Poço W4.

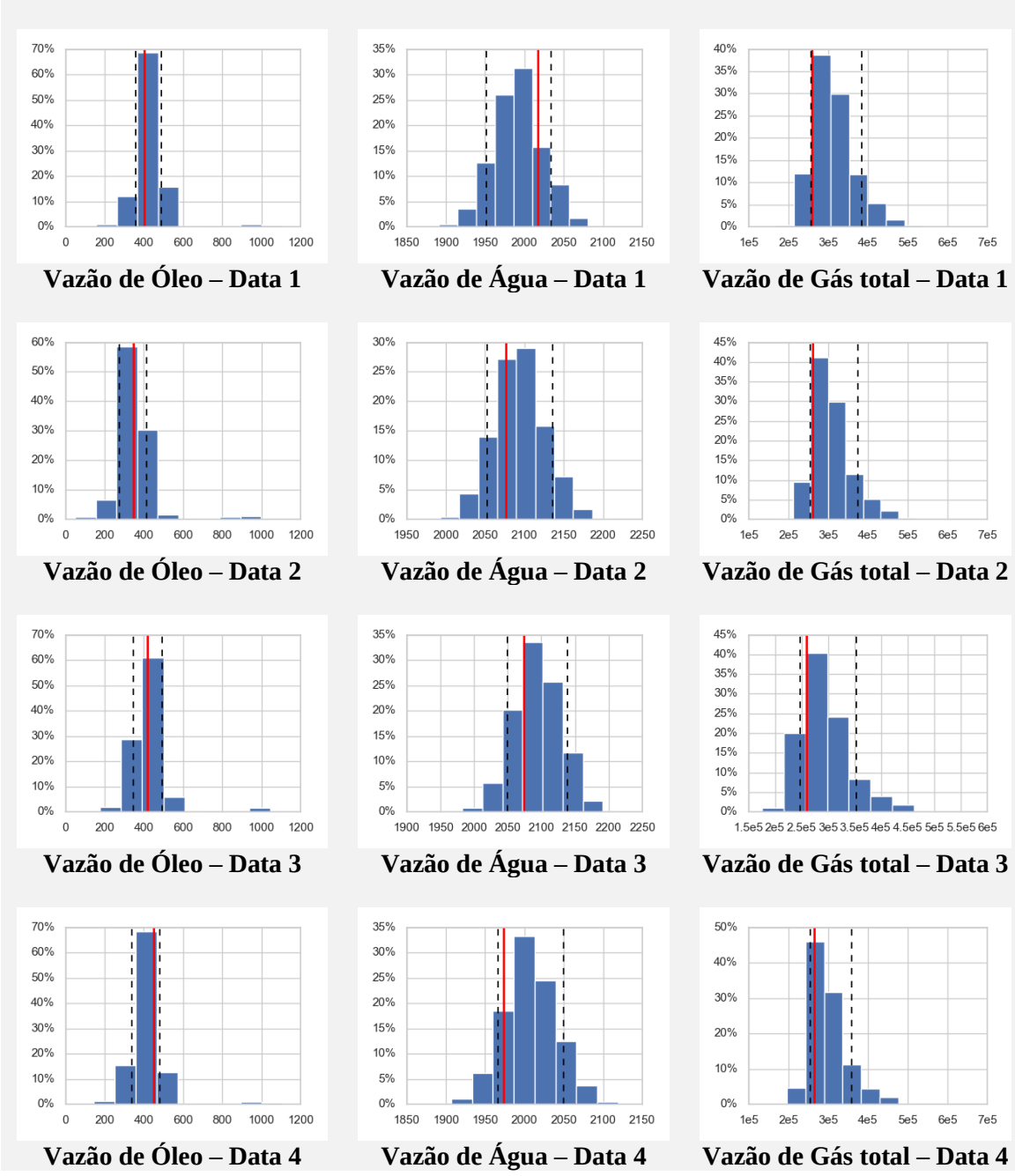


Figura 83: Resultados previstos para modelo MLR. Poço W5.

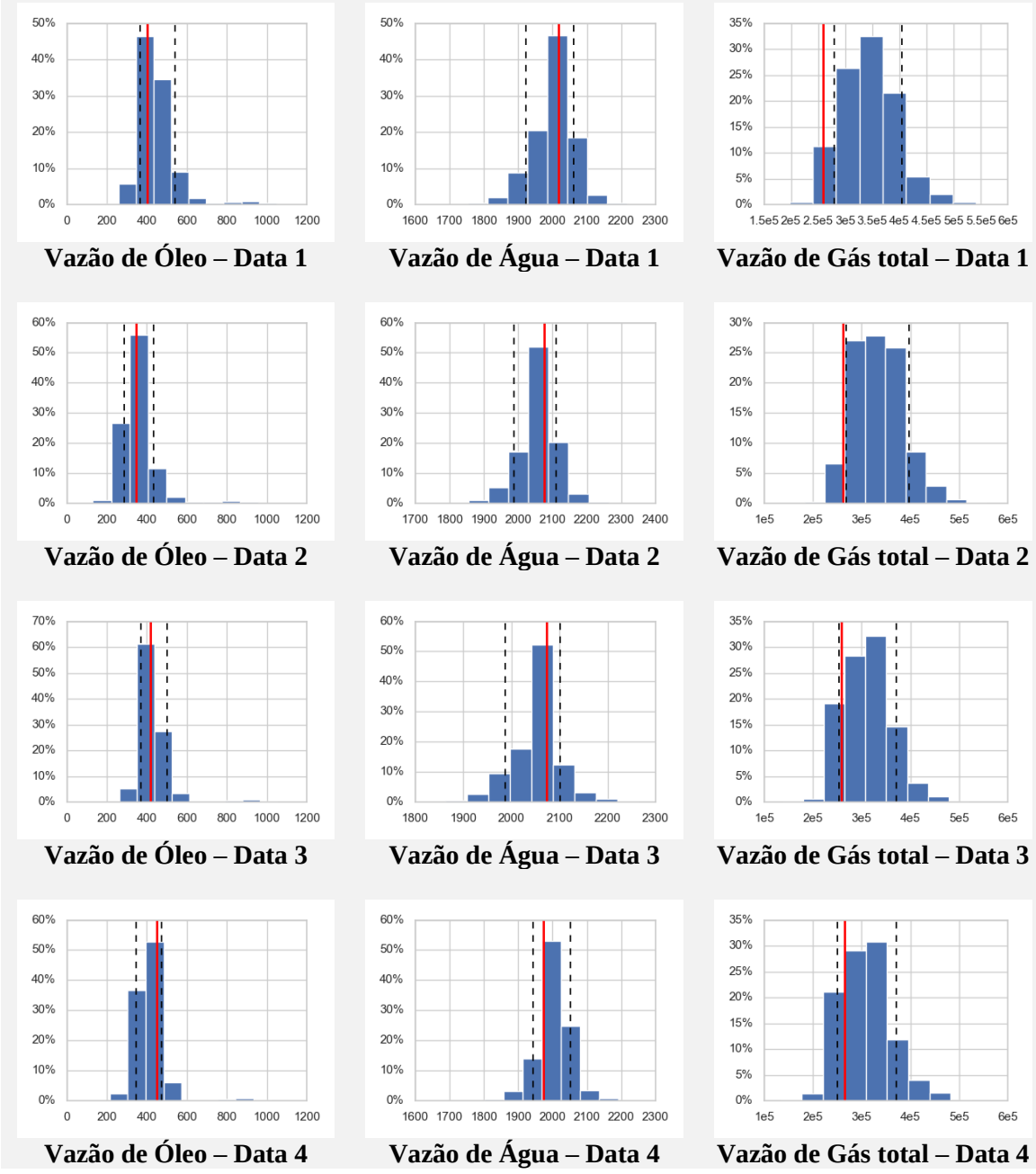


Figura 84: Resultados previstos para modelo SVR. Poço W5.

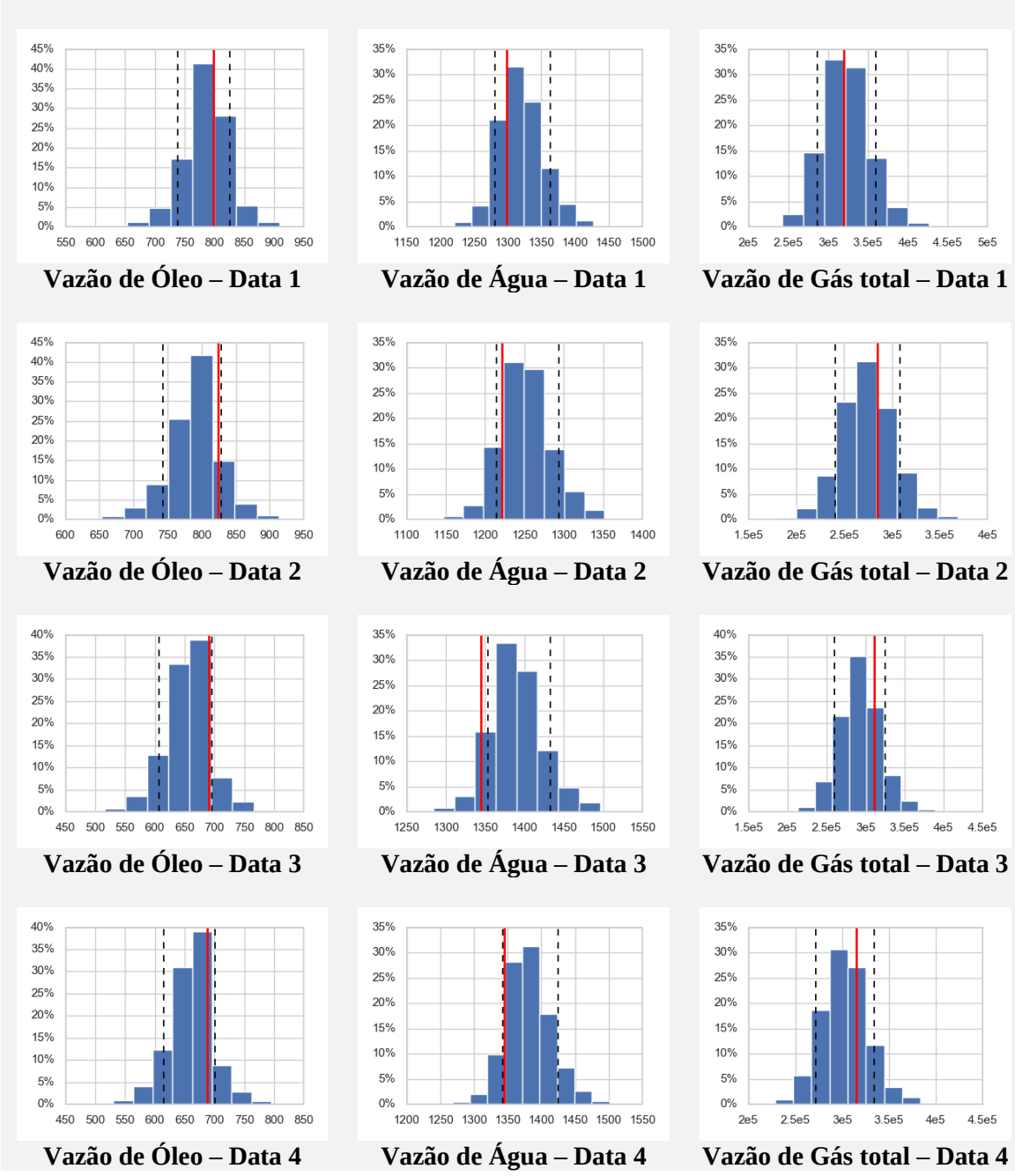


Figura 85: Resultados previstos para modelo MLR. Poço W6.

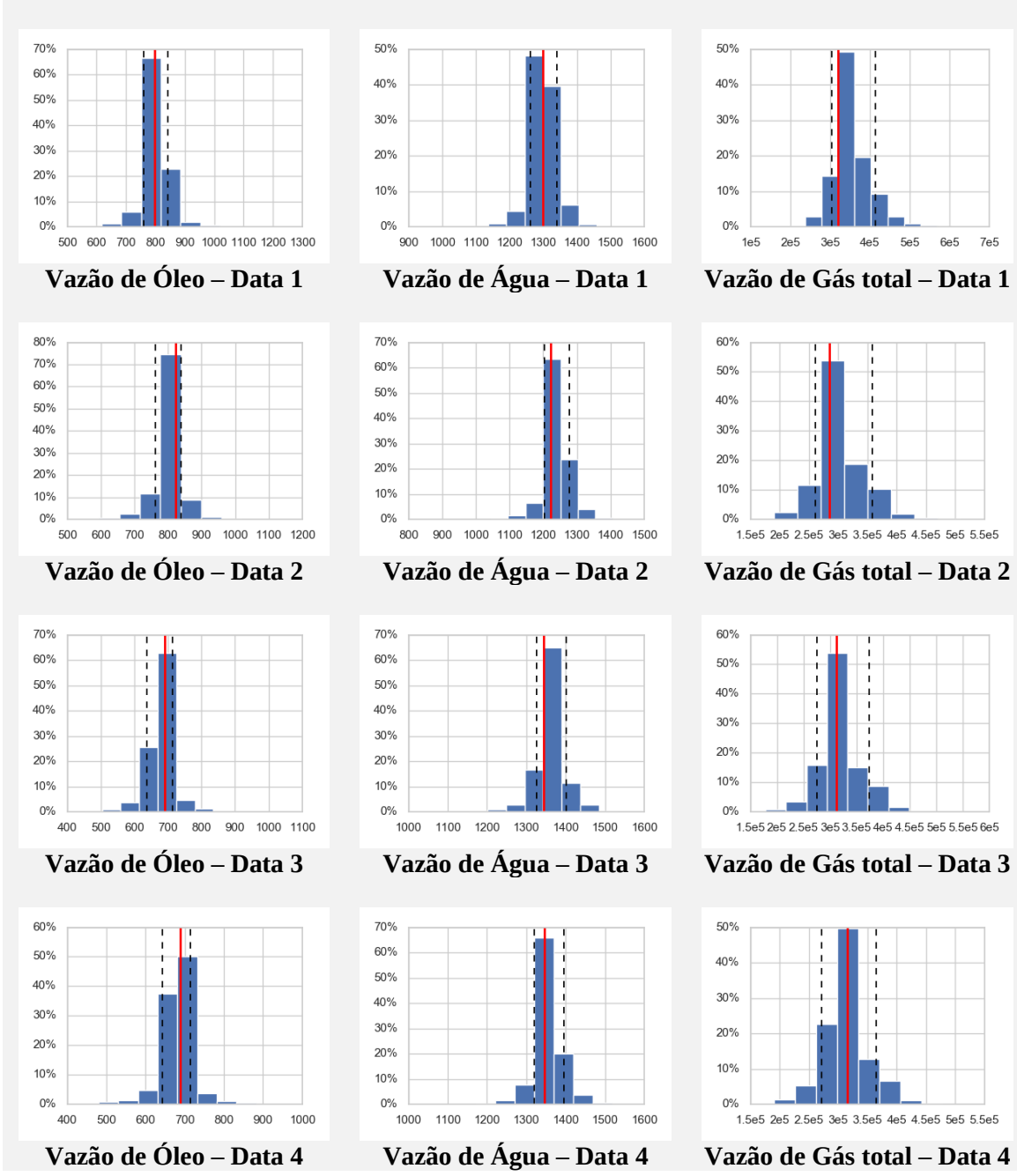


Figura 86: Resultados previstos para modelo SVR. Poço W6.

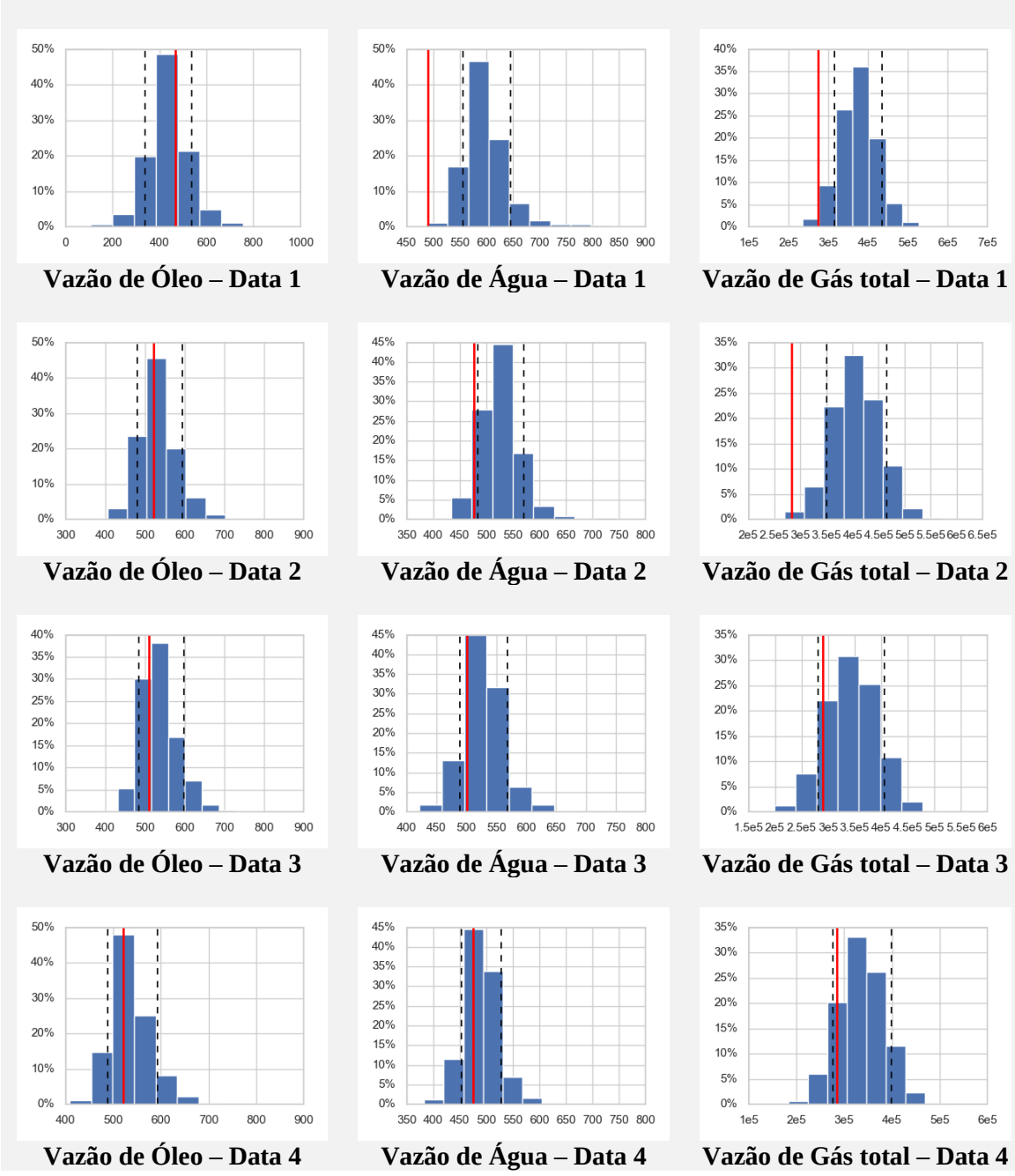


Figura 87: Resultados previstos para modelo MLR. Poço W7.

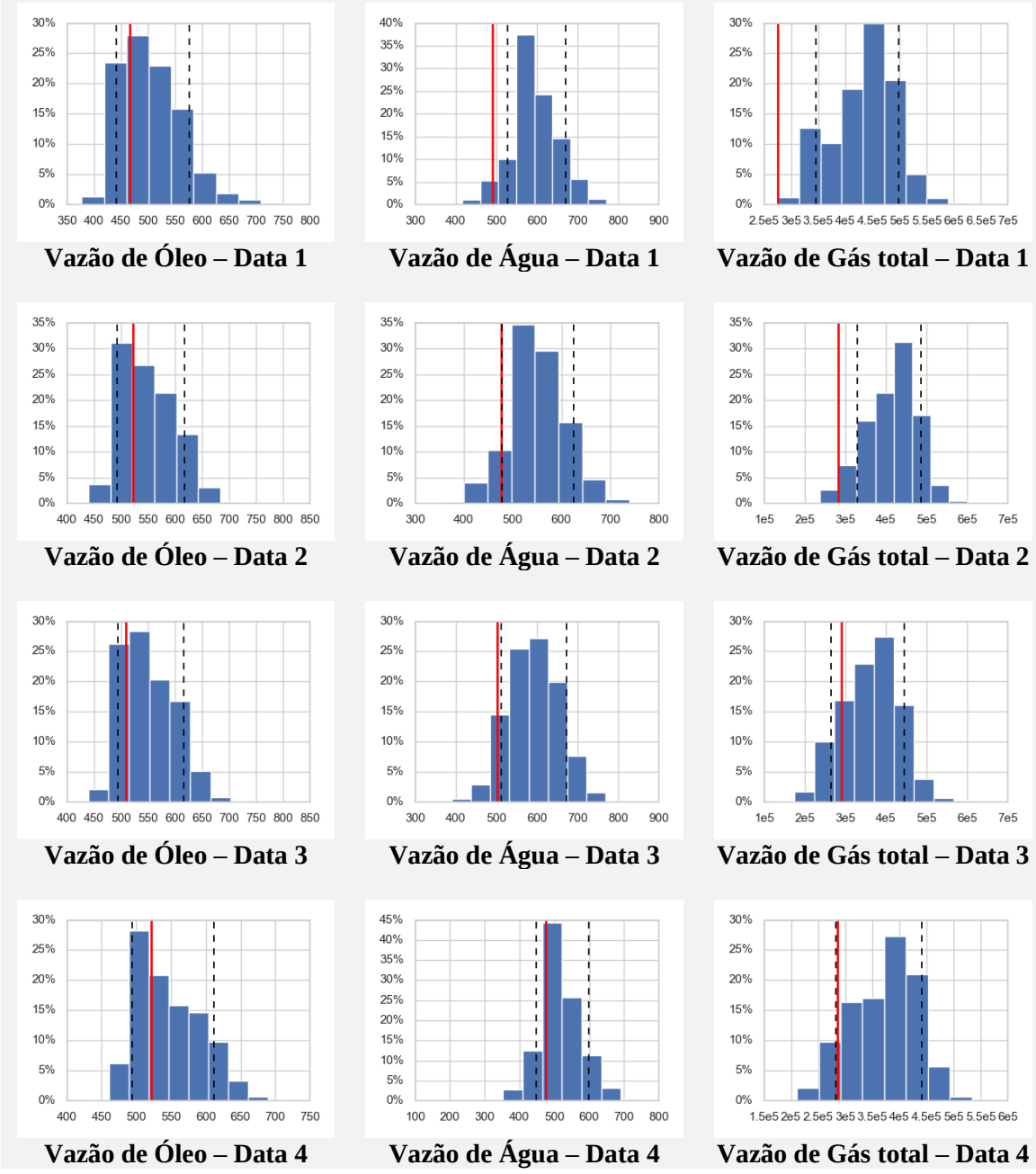


Figura 88: Resultados previstos para modelo SVR. Poço W7.

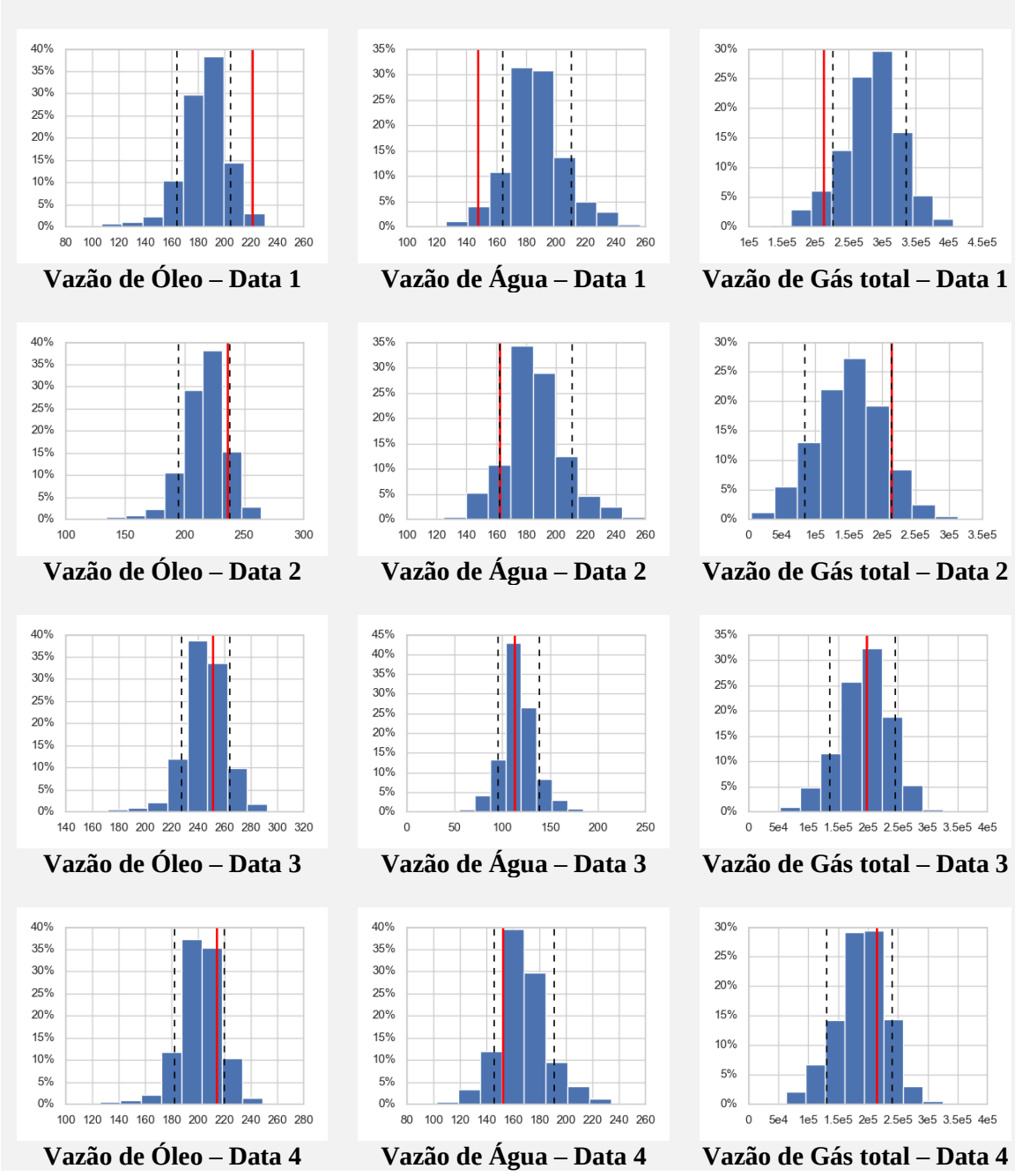


Figura 89: Resultados previstos para modelo MLR. Poço W8.

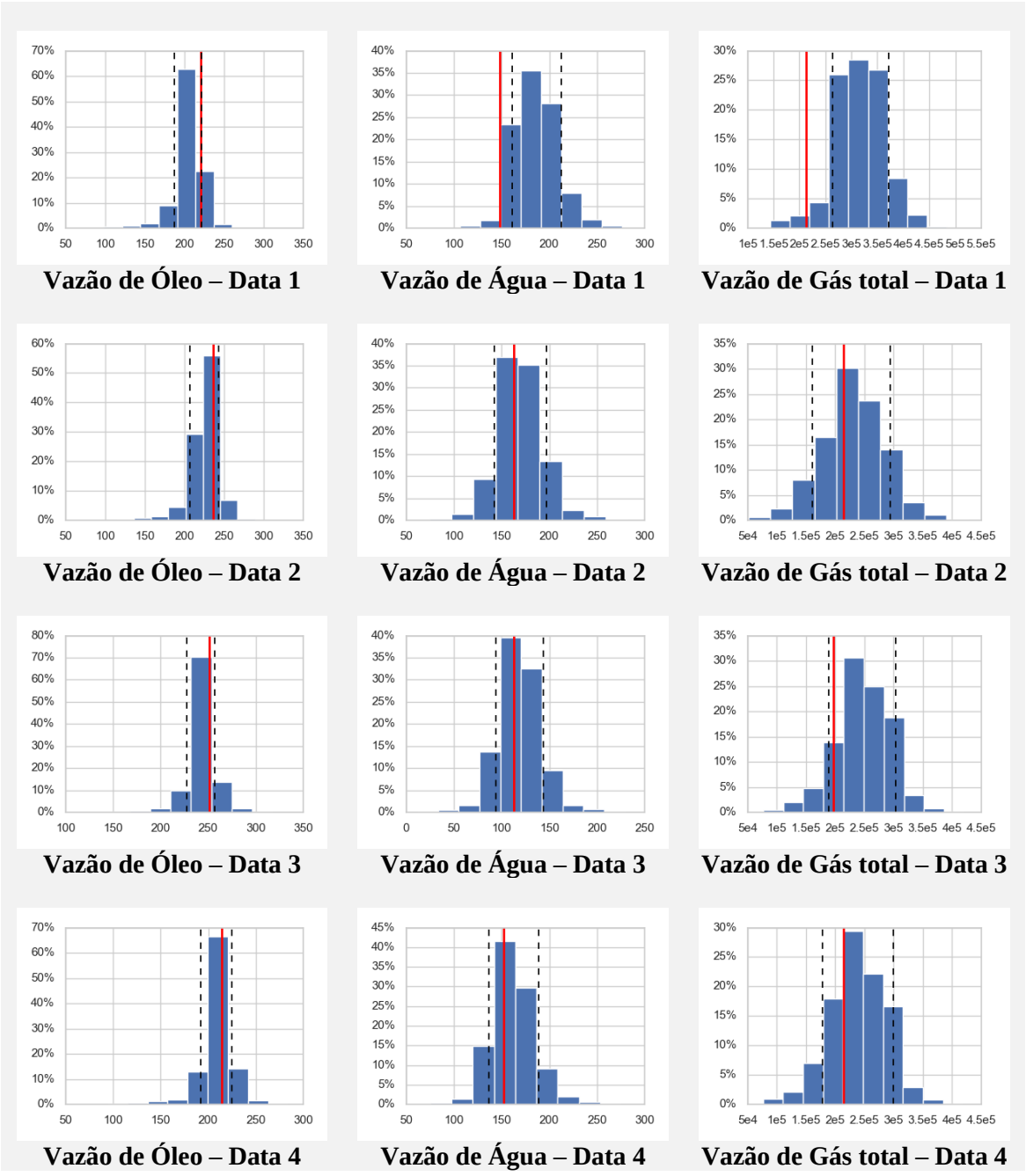
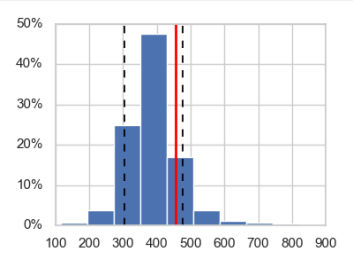
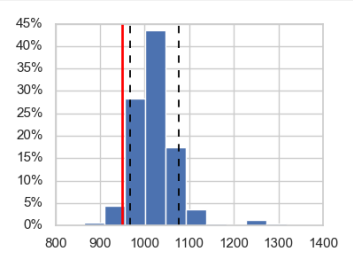


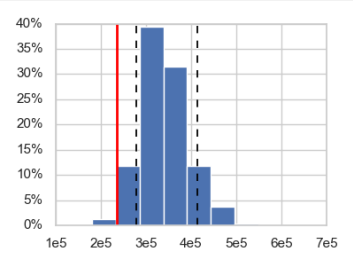
Figura 90: Resultados previstos para modelo SVR. Poço W8.



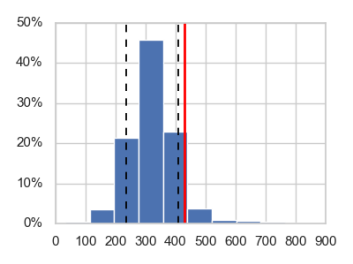
Vazão de Óleo – Data 1



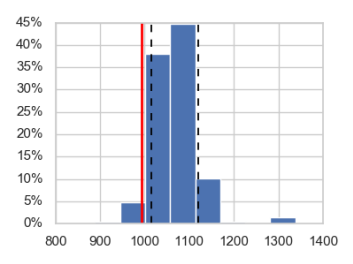
Vazão de Água – Data 1



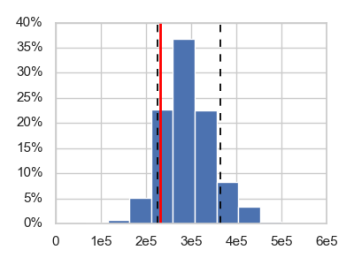
Vazão de Gás total – Data 1



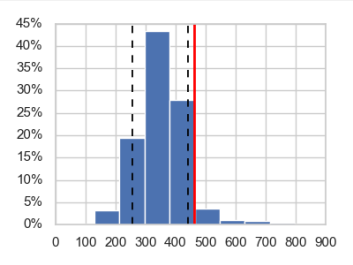
Vazão de Óleo – Data 2



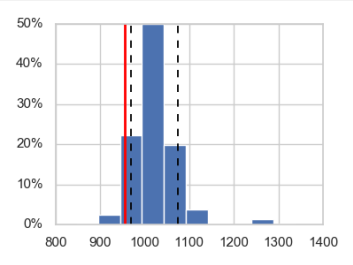
Vazão de Água – Data 2



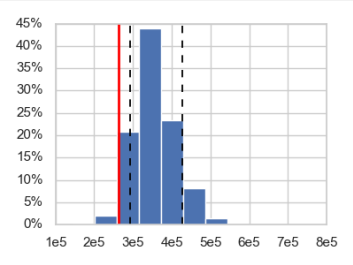
Vazão de Gás total – Data 2



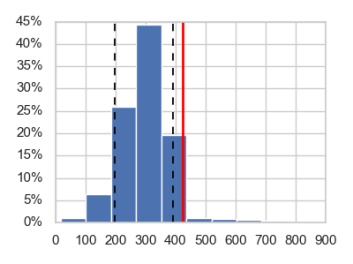
Vazão de Óleo – Data 3



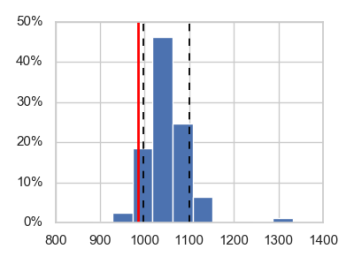
Vazão de Água – Data 3



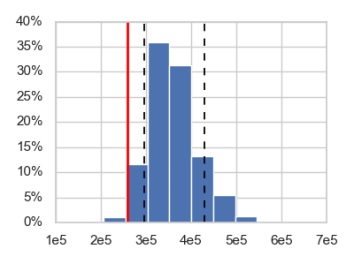
Vazão de Gás total – Data 3



Vazão de Óleo – Data 4



Vazão de Água – Data 4



Vazão de Gás total – Data 4

Figura 91: Resultados previstos para modelo MLR. Poço W9.

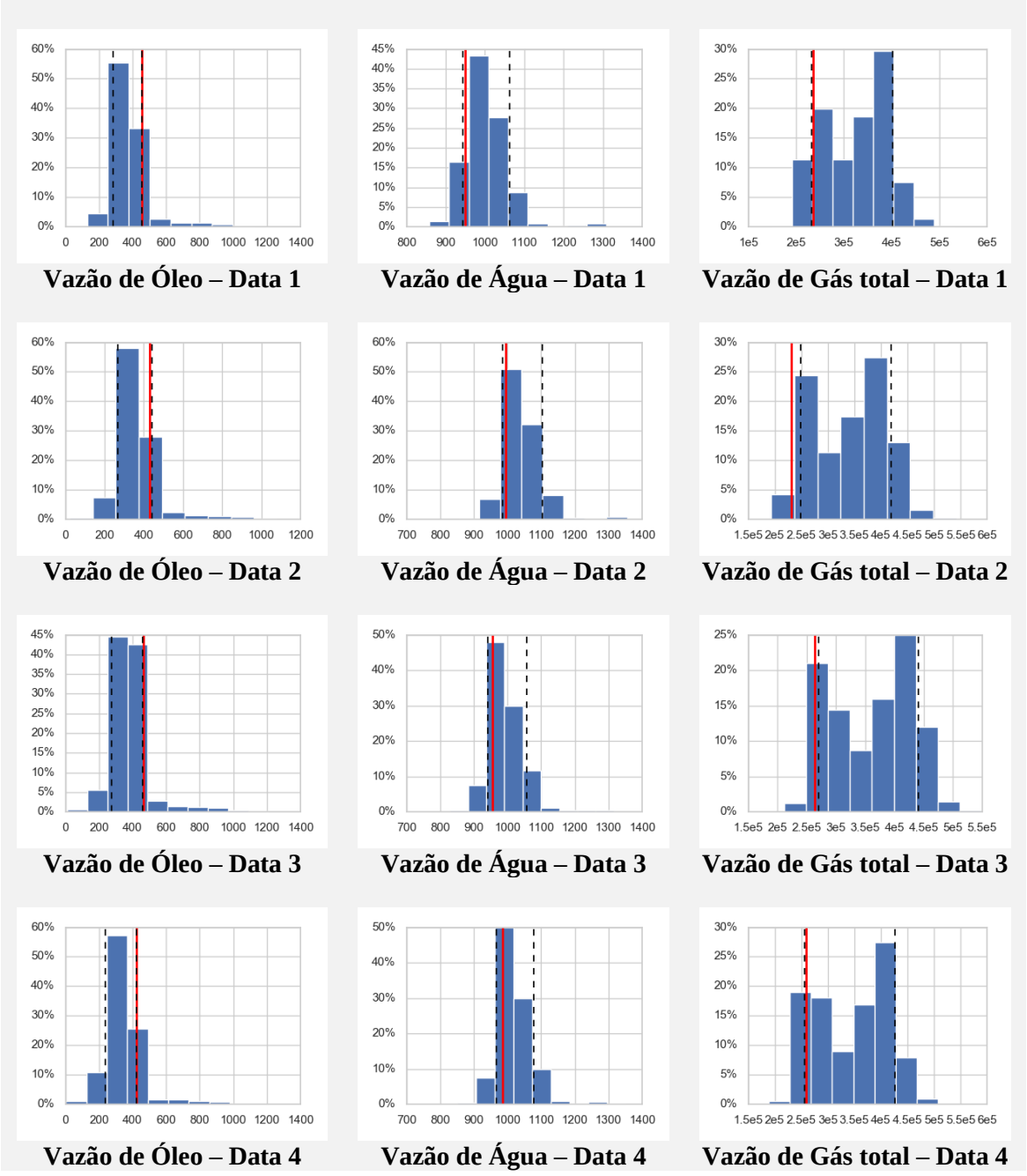


Figura 92: Resultados previstos para modelo SVR. Poço W9.

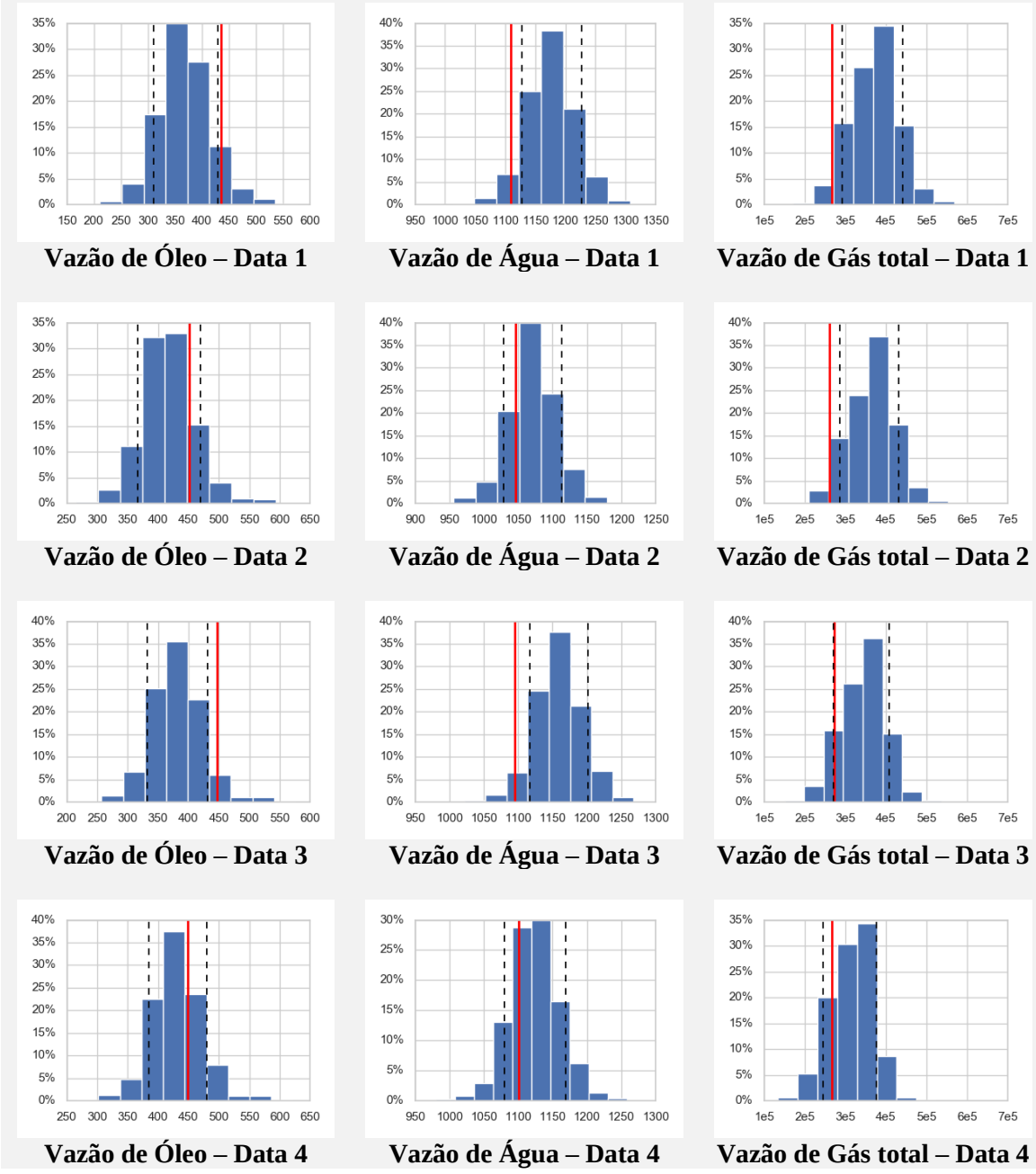


Figura 93: Resultados previstos para modelo MLR. Poço W10.

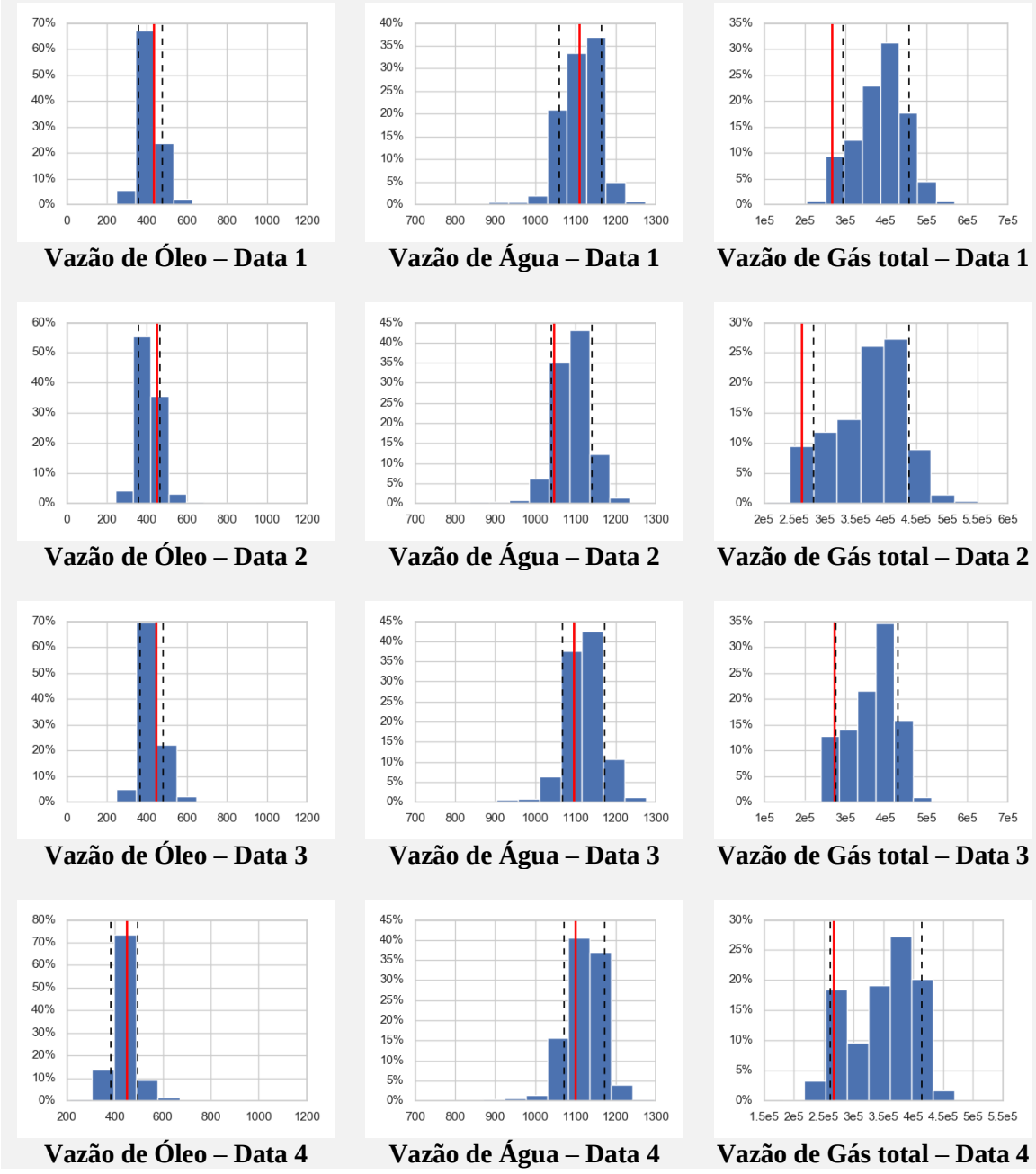


Figura 94: Resultados previstos para modelo SVR. Poço W10.

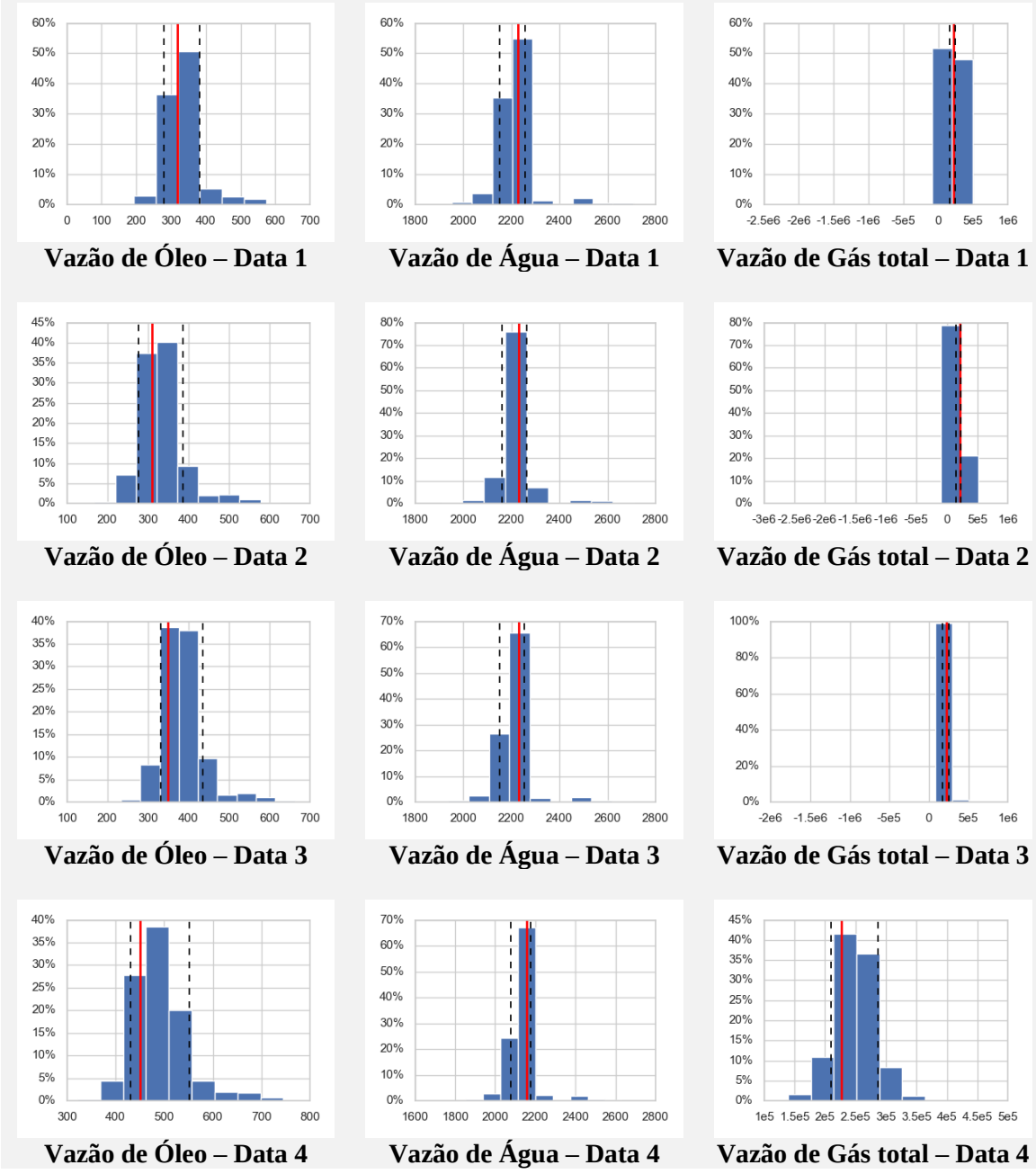
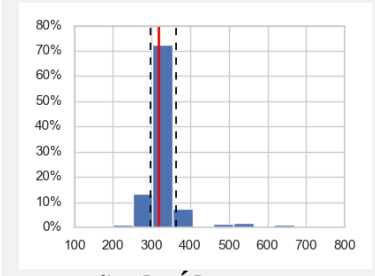
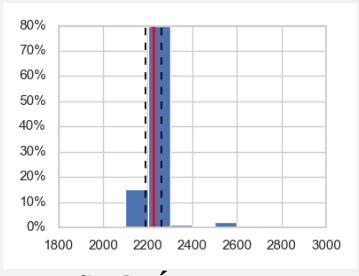


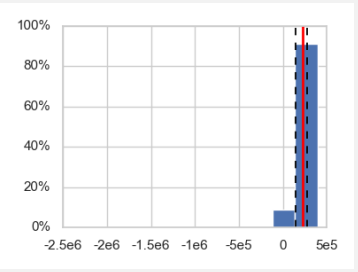
Figura 95: Resultados previstos para modelo MLR. Poço W11.



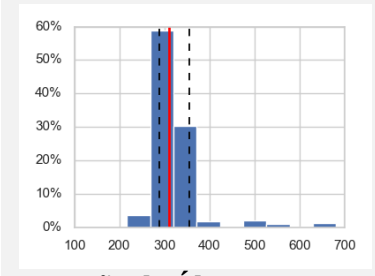
Vazão de Óleo – Data 1



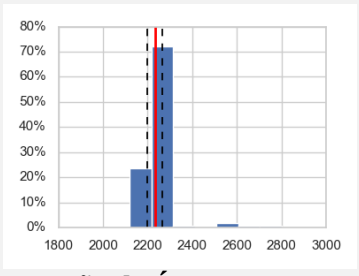
Vazão de Água – Data 1



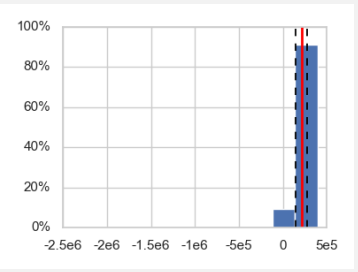
Vazão de Gás total – Data 1



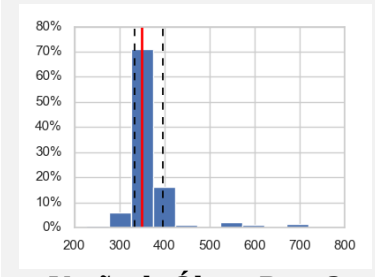
Vazão de Óleo – Data 2



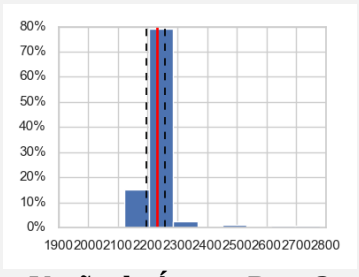
Vazão de Água – Data 2



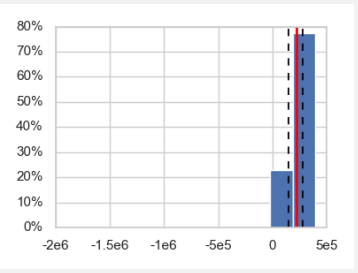
Vazão de Gás total – Data 2



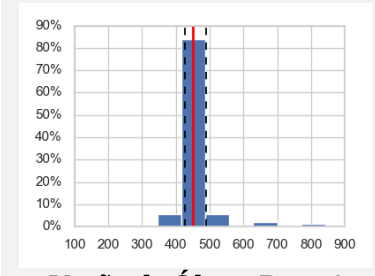
Vazão de Óleo – Data 3



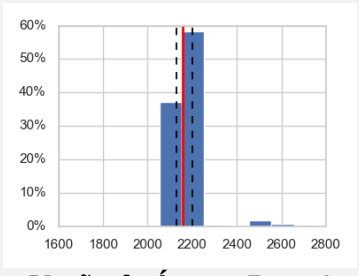
Vazão de Água – Data 3



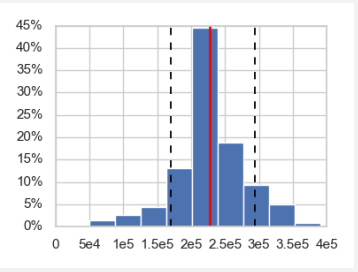
Vazão de Gás total – Data 3



Vazão de Óleo – Data 4

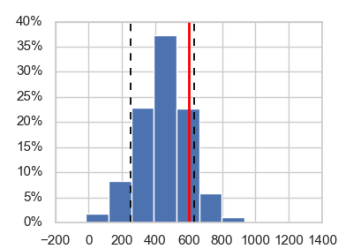


Vazão de Água – Data 4

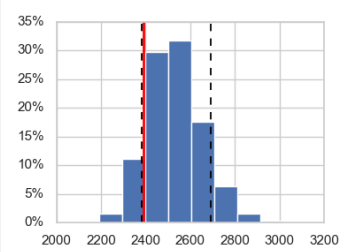


Vazão de Gás total – Data 4

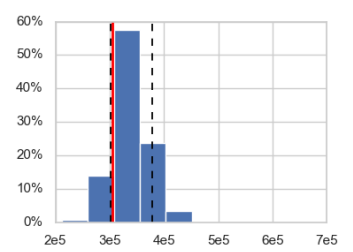
Figura 96: Resultados previstos para modelo SVR. Poço W11.



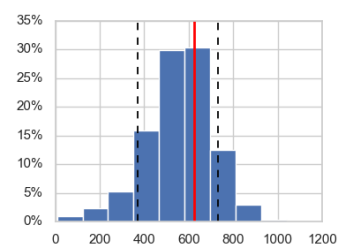
Vazão de Óleo – Data 1



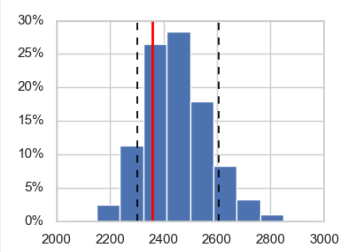
Vazão de Água – Data 1



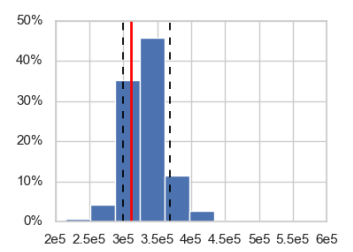
Vazão de Gás total – Data 1



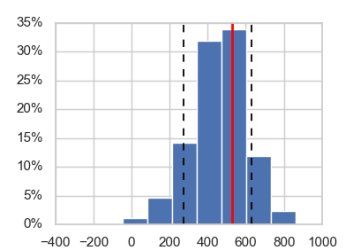
Vazão de Óleo – Data 2



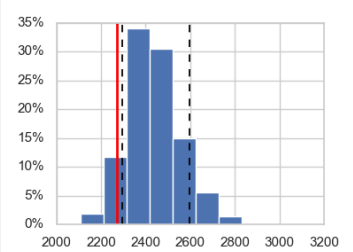
Vazão de Água – Data 2



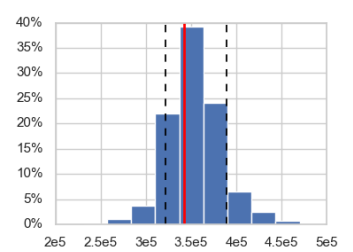
Vazão de Gás total – Data 2



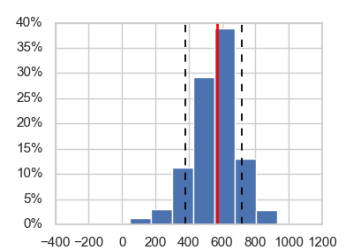
Vazão de Óleo – Data 3



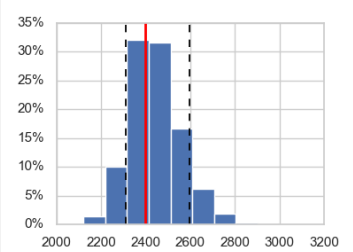
Vazão de Água – Data 3



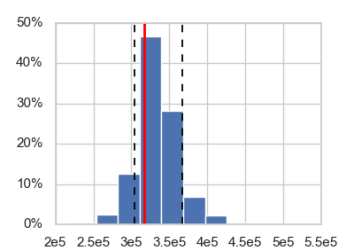
Vazão de Gás total – Data 3



Vazão de Óleo – Data 4



Vazão de Água – Data 4



Vazão de Gás total – Data 4

Figura 97: Resultados previstos para modelo MLR. Poço W12.

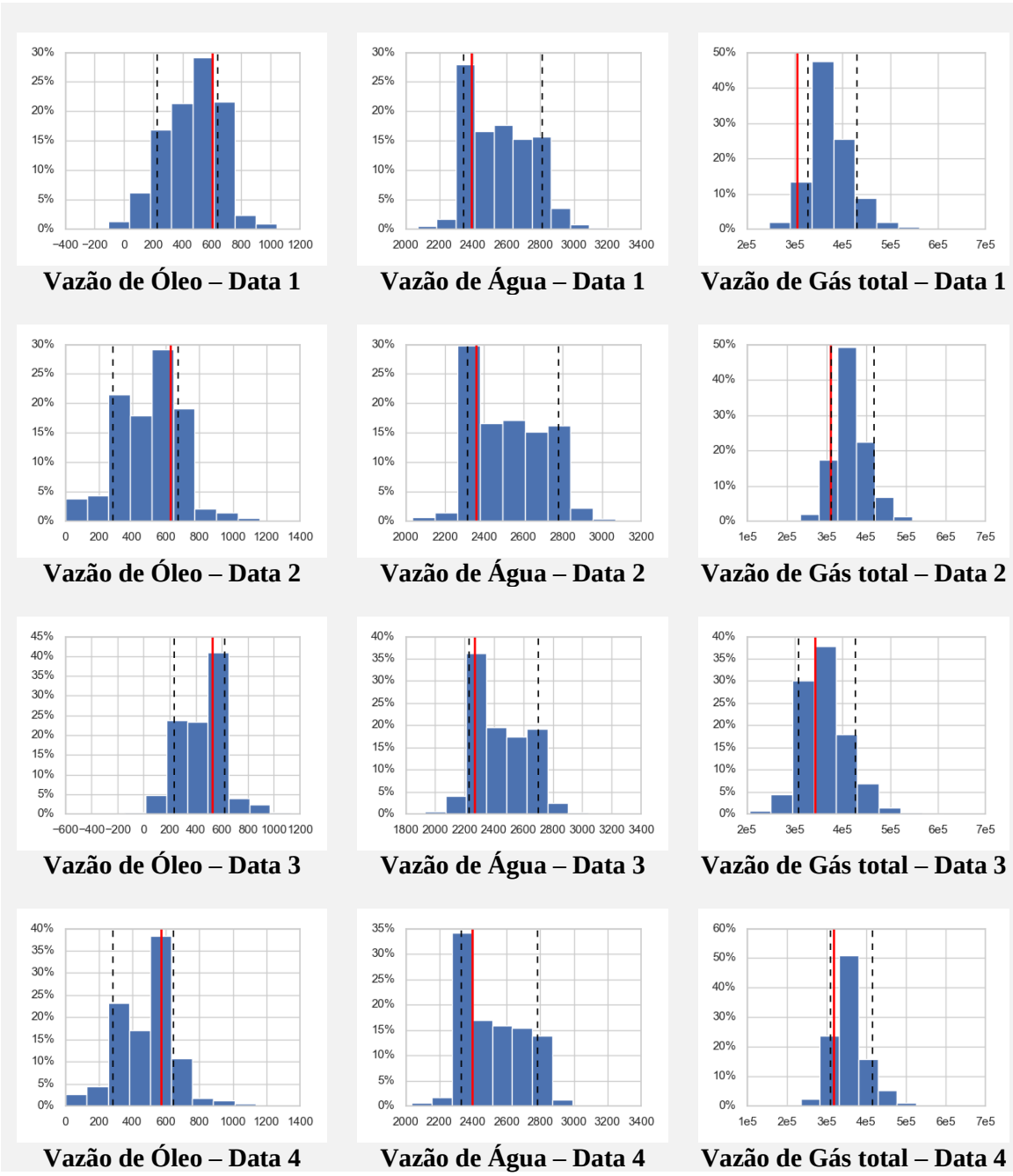


Figura 98: Resultados previstos para modelo SVR. Poço W12.

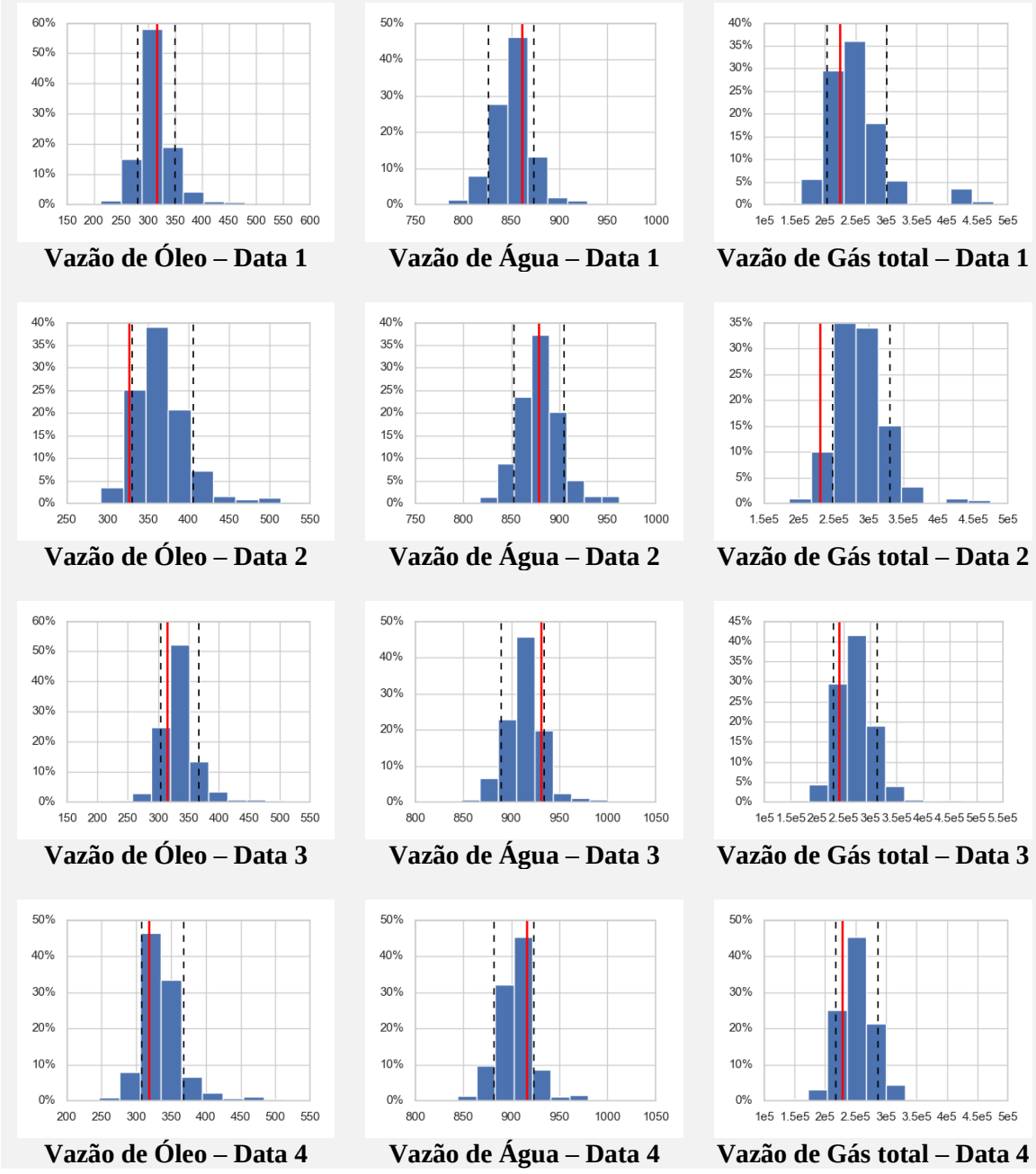


Figura 99: Resultados previstos para modelo MLR. Poço W13.

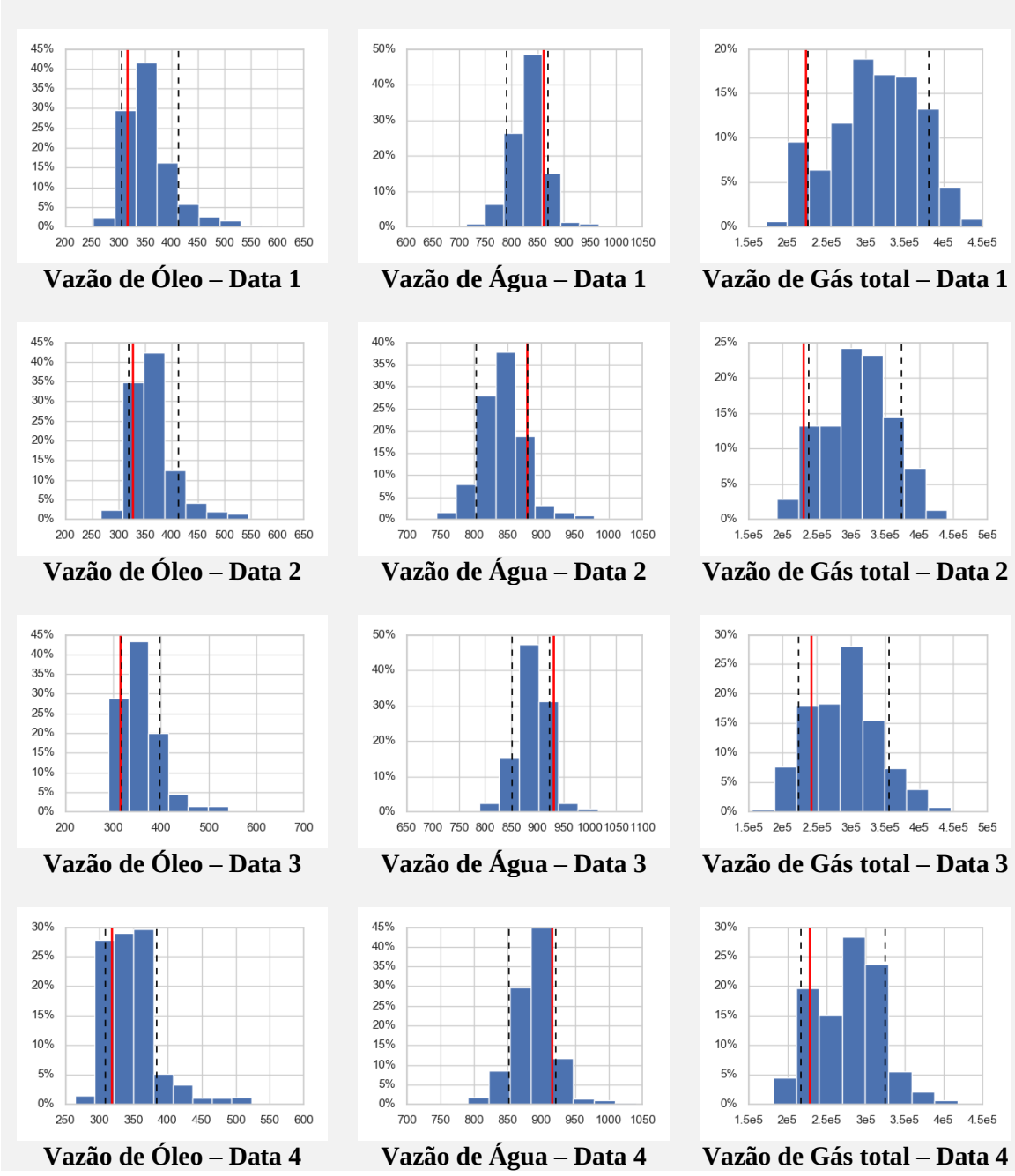


Figura 100: Resultados previstos para modelo SVR. Poço W13.