Mario Martins Ramos

Estratégias metodológicas associadas ao método *naïve* Bayes no problema da classificação de eletrofácies a partir de perfis geofísicos de poços: Aplicação no campo de Massapê, Bacia do Recôncavo, Bahia - Brasil

Niterói, RJ - Brasil

3 de setembro de 2022

Mario Martins Ramos

Estratégias metodológicas associadas ao método *naïve* Bayes no problema da classificação de eletrofácies a partir de perfis geofísicos de poços: Aplicação no campo de Massapê, Bacia do Recôncavo, Bahia - Brasil

Tese de doutorado apresentada ao programa de pós-graduação em Dinâmica dos Oceanos e da Terra (DOT) como exigência parcial para obtenção do título doutor em Geologia e Geofísica.

Universidade Federal Fluminense - UFF

Orientador: Rodrigo Bijani

Niterói, RJ - Brasil
3 de setembro de 2022

Mario Martins Ramos

Estratégias metodológicas associadas ao método *naïve* Bayes no problema da classificação de eletrofácies a partir de perfis geofísicos de poços: Aplicação no campo de Massapê, Bacia do Recôncavo, Bahia - Brasil

Tese de doutorado apresentada ao programa de pós-graduação em Dinâmica dos Oceanos e da Terra (DOT) como exigência parcial para obtenção do título doutor em Geologia e Geofísica.

Avaliado pela seguinte comissão examinadora:

Prof. Dr. Rodrigo Bijani
Orientador - PPGDOT/UFF

Prof. Dr. Cosme Ferreira Ponte-Neto
ON/MCTIC

Prof. Dr. Miguel Angelo Mane
PGG/UERJ

Prof. Dr. Arthur Ayres Neto
PPGDOT/UFF

Prof. Dr. Wagner Moreira Lupinacci
GIECAR-PPGDOT/UFF

Profª. Drª. Flora Ferreira Solon
PPGDOT/UFF

Dr. Diego Takahashi Tomazella
PPGEO - ON/MCTIC

Niterói, RJ - Brasil
3 de setembro de 2022

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

R175e Ramos, Mario Martins
Estratégias metodológicas associadas ao método naïve Bayes no problema da classificação de eletrofácies a partir de perfis geofísicos de poços: Aplicação no Campo de Massapê, Bacia do Recôncavo, Bahia - Brasil / Mario Martins Ramos ; Rodrigo Bijani, orientador. Niterói, 2022.
103 p. : il.

Tese (doutorado)-Universidade Federal Fluminense, Niterói, 2022.

DOI: http://dx.doi.org/10.22409/PPGDOT .2022.d.05724062747

1. Naïve Bayes. 2. Perfilagem Geofísica. 3. Bacia do Recôncavo. 4. Tuning. 5. Produção intelectual. I. Bijani, Rodrigo, orientador. II. Universidade Federal Fluminense. Instituto de Geociências. III. Título.

CDD -

Bibliotecário responsável: Debora do Nascimento - CRB7/6368

Agradecimentos

Eu agradeço primeiramente a todos os cientistas, a todos os curiosos, e aos insatisfeitos que vieram antes de mim pois, sem eles, nenhum trabalho não seria possível. Agradeço principalmente aos meus amigos e minha família, em particular ao meu Padrasto Sérgio Luiz Habib Pacca pela minha criação e por me ajudar e sempre me incentivar a seguir em frente; e à minha mãe Ana Luisa Vieira Martins por todo apoio que sempre me deu durante toda minha vida. Eu também agradeço, postumamente, a minha tia Maria Fernanda Vieira Martins, por todos esses anos de convívio, por sempre ajudar a mim e a minha mãe, e pelos valorosos ensinamentos durante todos esses anos.

Estendo meus agradecimentos aos professores, à secretaria, e aos demais funcionários do departamento de geologia e geofísica (GGO) da Universidade Federal Fluminense (UFF), por manterem a instituição sempre ativa e produtiva, mesmo no período da pandemia. Suas ações rápidas e eficientes certamente salvaram muitas vidas. Reservo também um espaço especial nesta tese para agradecer, como brasileiro, aos demais professores, cientistas, pesquisadores, e todos os profissionais envolvidos direta ou indiretamente na manutenção da ciência neste país, em especial àqueles responsáveis pelo desenvolvimento das vacinas contra a COVID-19. Agradeço às instituições de fomento pela manutenção das atividades mesmo neste período conturbado, especialmente pela CAPES, em relação a bolsa de pesquisa concedida no início do doutorado, que foi de fundamental relevância para a realização deste doutorado. Agradeço também aos estudantes da disciplina de Python (em particular nos anos de 2017, 2018 e 2019), e aos integrantes do grupo de pesquisa GIECAR por toda troca de ideias durante o meu doutorado. Agradeço à PETROBRAS pela bolsa de pesquisa concedida durante grande parte deste doutorado visto que este suporte financeiro foi fundamental para a realização deste trabalho, na aquisição de material didático, e no meu desenvolvimento pessoal. Adiciono a minha gratidão à empresa pelos cursos e palestras fornecidos pelos seus especialistas. Agradeço também a Eloíse Helena Policarpo Neves por me conduzir à leitura, algo que foi fundamental para melhorar minha escrita e vocabulário.

Em particular, sobre o conteúdo da tese, agradeço a Fernando Vizeu pelas valorosas discussões durante todo o período, assim como pelas sugestões da estimativa da densidade do kernel, e das métricas do valor-f e matriz de confusão; a Rodrigo Bijani e Rodrigo Dutra pela oportunidade de co-participar como orientador do trabalho que hoje considero como modelagem direta do *naïve* Bayes e da estratégia desenvolvida neste trabalho; à Wagner Moreira Lupinacci por todo suporte em relação a petrofísica, e pela sugestão da redução do número de fácies, o que contribuiu consideravelmente para o resultado da tese. Agradeço também a Antônio Fernando Menezes Freire pela disponibilização dos

dados deste trabalho, incluindo a interpretação litológica, e a disponibilização das zonas de produção em profundidade; e à Carolina Ferreira da Silva pela seleção, gerenciamento e processamento dos poços, que foi de fundamental importância para todo o grupo. Agradeço às valorosas contribuições dos membros externos da banca de qualificação de doutorado, que foram Cosme Ferreira da Ponte Neto, cujas sugestões de *bootstrap* contribuíram imensamente na redução do custo computacional nos testes dos códigos em Python; e Valéria Cristina Ferreira Barbosa cujas sugestões de leitura resultaram nas estratégias de *tuning* e arquitetura. Adicionalmente agradeço aos membros da Banca, em particular à Miguel Ângelo Mané, Arthur Ayres Neto, Flora Ferreira Solon e Diego Takahashi Tomazella por se disponibilizarem a avaliar este trabalho. Por fim agradeço ao meu orientador Rodrigo Bijani por me conduzir e orientar no meu doutorado, por me ajudar a formular os conceitos sobre os quais esta tese foi conduzida, e por acreditar em mim e no método naïve Bayes nos momentos em que nem eu mesmo acreditava.

Resumo

Este trabalho apresenta o potencial sistemático do método *naïve* Bayes, através da utilização de cinco estratégias de aprimoramento na classificação litológica a partir de dados de poços do Campo de Massapê, Bacia do Recôncavo. Existe uma quantidade considerável de trabalhos sobre os métodos de Aprendizado de Máquina, incluindo nas geociências, com grande enfoque na análise comparativa, mas com limitada discussão sobre a associação entre a metodologia e os dados. Usualmente, todos os método de aprendizado possuem um propósito metodológico atrelado à sua gênese, isto é, estes se propõem a resolver um problema específico, e é nesta especificidade que atingem a sua excelência. Portanto, o método *naïve* Bayes (NB) foi aplicado juntamente com quatro estratégias que visavam aprimorar o método. A primeira estratégia, denominada de KDE, em referência a Kernel Density Estimation tende a promover um melhor ajuste quanto à amostragem; a segunda utiliza uma técnica de aprimoramento denominada tuning, e foi denominada como TUN; a terceira está relacionada à uma modificação na estrutura, ou arquitetura de execução do NB padrão, e foi denominada de ARC; por fim, a última estratégia considera zonas estratigráficas em profundidade, e foi denominada de CRC. Foram utilizados os dados de 14 poços do Campo de Massapê (Bacia do Recôncavo), e os perfis GR, log(ILD) e DT, em turbiditos do Valanginiano (Eo-Cretáceo), considerado como principal reservatório da região. Também foram utilizadas as litologias categorizadas através da curva DRDN, que é referente a um método alternativo para a identificação de rochas siliciclásticas. Foram feitas duas análises separadas que foram: As validações com dados reais e uma classificação de rochas em um único poço sem perfil litológico. Os resultados da validação indicaram que todas as estratégias promovem um ganho incipiente em relação ao NB, comprovando a reconhecida estabilidade (e robustez) deste classificador Bayesiano. Já a classificação demonstrou grande variedade de resultados das estratégias KDE e ARC em relação às demais.

Palavras-chave: Naïve Bayes, Tuning, Committee, Bacia do Recôncavo, Perfilagem Geofísica, Formação Maracangalha, Membro Caruaçu.

Abstract

This work presents the systematic potential of the *naïve* Bayes method, through the use of four improvement strategies in lithological classification based on well data from Massapê Field, Recôncavo Basin. There is a considerable amount of work on Machine Learning methods, including in the geosciences, with a strong focus on comparative analysis, and limited discussion on the association between methodology and data. Usually, all learning methods have a methodological purpose linked to their genesis, in other words, they propose to solve a specific problem, and it is in this specificity that they reach their excellence. Therefore, the *naïve* Bayes (NB) method was applied together with four strategies that aimed to improve the method. The first strategy, called KDE, in reference to Kernel Density Estimation tends to promote a better fit in terms of sampling; the second uses an improvement technique called tuning, and was called TUN; the third is related to a modification in the structure, or execution architecture of the standard NB, and was called ARC; finally, the last strategy considers stratigraphic zones in depth, and was called CRC. Data from 14 wells in the Massapê Field (Recôncavo Basin) and the GR, log(ILD) and DT profiles were used in Valanginian (Eo-Cretaceous) turbidites, considered the main reservoir in the region. Lithologies categorized through the DRDN curve were also used, which refers to an alternative method for the identification of siliciclastic rocks. Two separate analyzes were performed which were: Validations with real data and a rock classification in a single well without lithological profile. The validation results indicated that all strategies promote an incipient gain in relation to NB, proving the recognized stability (and robustness) of this Bayesian classifier. The classification showed a great variety of results of the KDE and ARC strategies in relation to the others.

Keywords— Naive Bayes, Tuning, Committee, Recôncavo Basin, Well logging, Maracangalha Formation, Caruaçu Member

Sumário

Sumário 9

Lista de ilustrações 12

1 INTRODUÇÃO 16
1.1 Motivação científica 19
1.2 Aspectos Inovadores 22

2 GEOLOGIA REGIONAL 23
2.1 Estruturação e Embasamento 24
2.2 Evolução Sedimentar 24
2.3 Formação Maracangalha e Membros Caruaçu e Pitangas 26
2.4 Perfis Geofísicos de Poços 28

3 METODOLOGIA 29
3.1 Dados de Poços 30
3.2 Banco de Dados de Treinamento 32
3.3 Métricas de Validação 33
3.3.1 Perfil Erros 33
3.3.2 Tipos de Erros 33
3.3.3 Matriz de Confusão 34
3.3.4 Precisão Revocação e valor F 35

4 TEOREMA DE BAYES 37
4.1 Priori 37
4.2 Verossimilhança 38
4.3 Regra de Bayes 38
4.4 FDP - Função de Densidade de Probabilidade 39
4.5 Condição *naïve* 39
4.6 *Naïve* Bayes padrão (STD) 41
4.7 Estratégias para Aprimorar o Método *Naïve* Bayes 42
4.8 1ª estratégia: Cálculo das verossimilhanças via *Kernel Density Estimation* (KDE) 42
4.9 2ª estratégia: Utilização do *Naïve* Bayes em Modo *ensemble* (Arquitetura - ARC) 43
4.10 3ª estratégia: *Naïve* Bayes Sintonizado (TUN) 44
4.10.1 Sequência de *Tuning* 45

4.11 4ª estratégia: *Naïve* Bayes Combinado com Informação Geológica Prévia (CRC) . . . 48

5 RESULTADOS . . . 49
5.1 Distribuição e Análise dos Dados de Treinamento para o Campo de Massapê . . . 49
5.1.1 Análise do Campo de Massapê: Erro e valor-f . . . 50
5.2 Poço de Validação :7-MP-50D-BA . . . 52
5.2.1 Avaliação da Classificação via STD . . . 52
5.2.2 Avaliação da Classificação via KDE . . . 55
5.2.3 Avaliação da Classificação via TUN . . . 57
5.2.4 Avaliação da Classificação via ARC . . . 59
5.2.5 Avaliação da classificação via CRC . . . 61
5.2.6 Poço de classificação . . . 64

6 CONCLUSÕES . . . 66

7 APÊNDICE . . . 68
7.1 Introduction . . . 68
7.2 Geologic settings: Recôncavo Basin . . . 71
7.2.1 Massapê Field . . . 72
7.3 Well-log data of Massapê Field . . . 73
7.3.1 Interpreted Lithologic Log: DRDN method . . . 73
7.4 Methodology . . . 74
7.4.1 Basic statistics . . . 74
7.4.2 Naïve-Bayes Classifier . . . 76
7.4.3 Probability Density estimation . . . 76
7.4.3.1 Normal distribution . . . 77
7.4.3.2 Kernel Density Estimation . . . 77
7.4.4 Standard naïve Bayes (NB) classifier . . . 79
7.4.5 Strategy 1: likelihoods using the Gaussian Kernel Density Estimation . . . 79
7.4.6 Strategy 2: Automatic definition of priors - Tuning . . . 79
7.4.7 Strategy 3: Altering Training data-sets, priors and likelihoods - Architecture 80
7.4.8 Strategy 4: Using stratigraphic information in depth-zones . . . 84
7.5 Results . . . 85
7.5.1 Selecting the validation well . . . 85
7.5.2 Validation well: 7MP-50D-BA . . . 86
7.5.3 Classification of well 7MP-50D-BA using the CRC strategy . . . 88
7.6 Discussions and Conclusions . . . 90
7.7 Acknowledgments . . . 92

REFERÊNCIAS 93

Lista de ilustrações

Figura 1 – Localização da Bacia do Recôncavo em laranja e suas feições estruturais principais (BRUHN, 1999). 23
Figura 2 – A - Seção da Bacia do Recôncavo esquemática (MAGNAVITA; SILVA, 1995); B - estimativa de profundidade da bacia por modelagem gravimétrica (SALES; BIJANI; FREIRE, 2019). 25
Figura 3 – Carta estratigráfica adaptada da Bacia do Recôncavo segundo Olívio et al. (2007). 26
Figura 4 – Localização dos poços no Campo de Massapê 28
Figura 5 – Banco de dados usado considerando os 12 poços de validação do Campo de Massapê. A litologia está apresentada na forma numérica onde 57 é o código pro folhelho; 49 para o arenito; e 25,0 para o *slurry*. As zonas também estão em codificação numérica onde: 111 é referente a CRC 1; 222 é referente a CRC 2; e 333 é referente a CRC 3. 32
Figura 6 – Exemplo esquemático do cálculo do perfil de erros. O primeiro perfil (à esquerda) na figura é o perfil litológico verdadeiro; o segundo (centro) consiste das litofácies classificadas; o terceiro (à direita) corresponde aos erros entre o verdadeiro e o modelado. 33
Figura 7 – Exemplo esquemático que apresenta os quatro tipos de erros: *True Negative* indica que a negativa em relação a um tipo de classificação foi correta. *False negative* indica erro na negativa em relação a classe vigente. *True positive* ocorre quando há acerto na classificação da classe e *False positive* está relacionado à afirmação em relação a uma classe for incorreta. 34
Figura 8 – Exemplo ilustrativo da matriz de confusão referente às duas litologias principais do campo de Massapê. 35
Figura 9 – Exemplo esquemático da *Precision* (Precisão), *Recall* (Revocação) e *fscore* (valor-f). 36
Figura 10 – Seção da Bacia do Recôncavo, onde a maior parte dos sedimentos é folhelho (verde), mas também com alguma quantidade considerável de arenito (amarelo) (Da Silva, 2013). 37
Figura 11 – Probabildiades baseadas em histogramas 2D (A,B e C) onde ρ é o coeficiente de correlação; e probabilidades segundo condição *naïve* (D, E e F). 40
Figura 12 – Exemplo de classificação segundo a estratégia ARC, onde w_{1-5} são os poços e L' a litologia classificada. O $*$ na linha 5 indica um valor de L' obtido pelo critério de desempate. 44

Figura 13 – Convergência das *prioris* considerando M=10 iterações. 47
Figura 14 – Gráficos de dispersão e distribuição do banco de dados de treinamento; Letras a, e, i apresentam as verossimilhanças por litologia, tanto ajustada via distribuição normal (curva nas cores da litologia), assim como valores da média $\bar{m}$ e do desvio padrão σ para cada litologia; a triangulação esquerda, referente as letras d, g, h, são os gráficos de dispersão para cada litologia, apresentando também o coeficiente de correlação r e gaussianas bidimensionais para cada litologia (curvas de nível nas cores da litologia); A triangulação direita, referente as letras b, c, f, apresenta o gráfico de compartimentação hexagonal, e as gaussianas bidimensionais para cada litologia com sua respectiva coloração. 50
Figura 15 – Tabela com os valores dos erros para cada poço, acrescido da média para cada estratégia. 51
Figura 16 – Tabela com os valores de MVF para cada poço, acrescido da média para cada estratégia. 51
Figura 17 – Gráficos em pizza para todos os 12 poços do Campo de Massapê. . . . 52
Figura 18 – Classificação do poço de validação via estratégia STD onde: (a) litologia interpretada; (b) perfil GR; (c) perfil DT; (d) perfil log(ILD); (e) litologia classificada; (f) perfil de erros; (g) perfil de Probabilidades. 54
Figura 19 – Matriz de confusão (a) e tabela com valores de precisão, revocação e valor-f (b) para o *naïve* Bayes STD 55
Figura 20 – Classificação do poço de validação via estratégia KDE, onde: (a) litologia interpretada; (b) perfil GR; (c) perfil DT; (d) perfil log(ILD); (e) litologia classificada; (f) perfil de erros; (g) perfil de Probabilidades. 56
Figura 21 – Matriz de confusão (a) e tabela com valores de precisão, revocação e valor-f (b) para a estratégia KDE 57
Figura 22 – *Priori* ótima calculada após 10 iterações. 57
Figura 23 – Classificação do poço de validação via estratégia TUN, onde: (a) litologia interpretada; (b) perfil GR; (c) perfil DT; (d) perfil log(ILD); (e) litologia classificada; (f) perfil de erros; (g) perfil de Probabilidades. 58
Figura 24 – Matriz de confusão (a) e tabela com valores de precisão, revocação e valor-f (b) para a estratégia TUN 59
Figura 25 – Classificação do poço de validação via estratégia ARC, onde: (a) litologia interpretada; (b) perfil GR; (c) perfil DT; (d) perfil log(ILD); (e) litologia classificada; (f) perfil de erros; (g) perfil de Probabilidades. 60
Figura 26 – Matriz de confusão (a) e tabela com valores de precisão, revocação e valor-f (b) para a estratégia ARC 61
Figura 27 – Gráficos em pizza mostrando as prioris para cada zona considerando o poço de validação 7-MP-50D-BA. 61

Figura 28 – Classificação do poço de validação via estratégia CRC, onde: (a) litologia interpretada; (b) perfil GR; (c) perfil DT; (d) perfil log(ILD); (e) litologia classificada; (f) perfil de erros; (g) perfil de Probabilidades. 62
Figura 29 – Matriz de confusão (a) e tabela com valores de precisão, revocação e valor-f (b) para a estratégia KDE 63
Figura 30 – Classificações do poço de validação usando todas as estratégias, onde: (a) litologia interpretada; (b) perfil GR; (c) perfil DT; (d) perfil log(ILD); (e) estratégia STD; (f) estratégia KDE; (g) estratégia CRC; (h) estratégia ARC; (i) estratégia TUN; 64
Figura 31 – Aplicação das estratégias ao poço de classificação 5-BRSA-365-BA, onde: (a) perfil GR; (b) perfil DT; (c) perfil log(ILD); (d) Litologia classificada através da estratégia *naïve* Bayes; (e) Litologia classificada através da estratégia KDE; (f) Litologia classificada através da estratégia *naïve* Bayes faciológico; (g) Litologia classificada através da estratégia *naïve* Bayes arquitetura; (h) Litologia classificada através da estratégia *naïve* Bayes tuning 65
Figura 32 – (A) Location of the Recôncavo Basin in Brazil. (B) the study area comprising the Massapê oilfield. (C) The set of training wells (black dots) and the validation well (green dot) (BRUHN, 1999). 73
Figura 33 – Sketch of the concepts for true and false positives and negatives for a specific lithology (i.e., shale). 75
Figura 34 – Likelihoods computed by normal distribution (a, b, c) and using the Gaussian kernel density estimation (d, e, f) for each lithologies of the entire training data-set. 78
Figura 35 – (a) Convergence of priors using M=10 iterations and (b) the best obtained prior. 80
Figura 36 – Bar graph with all priors obtained for each training well. 81
Figura 37 – The worst (a, b and c) and the best (d, e and f) likelihood models obtained during the architecture procedure. $\bar{m}$ and σ are means and standard deviations for sandstones (yellow), shales (green) and slurries (grey). 82
Figura 38 – All classifications of the validation well 7-MP-50D-BA (a-l) obtained during the architecture procedure. Track m shows the final classification outcome. 83
Figura 39 – Pie plots showing different prior probabilities based on depth-zones for validation-well 7-MP-50D-BA. 84
Figura 40 – Flowchart of the proposed method. 85
Figura 41 – Error analysis by using the fscore calculation (i.e., Equation 7.4). . . . 86

Figura 42 – Overall classifications obtained for 7MP-50D-BA. (a) Interpreted Lithologic Log, (b) GR log, (c)DT log, (d) log(ILD) log. The classification outcome using (e) standard NB, (f) KDE, (g) CRC, (h) ARC, (i) TUN. 87
Figura 43 – Results for validation well 7MP-50D-BA. (a) Interpreted Lithologic Log, (b) GR log, (c) DT log, (d) log(ILD) log, (e) classification log, (f) error log, (g) probability log, (h) depth-zones. The zone CRC-1 varies from 2340 to 2495 m depth, while zone CRC-2 varies from 2495 to 2657. The CRC-3 zone goes from 2657 down to 2973 m depth. 89
Figura 44 – (a) Confusion Matrix and (b) Precision, recall and fscore of CRC (i.e., Equations 7.2, 7.3 and 7.4, respectively) strategy applied to the validation well 7MP-50D-BA. 90

1 Introdução

O aprendizado de máquina (AM) (*Machine Learning*, em inglês) é uma subárea da inteligência artificial que atribui a capacidade de aprendizado através dos dados a um computador[1] (HURWITZ; KIRSCH, 2018). O conceito de uma máquina (ou computador) capaz de aprender foi originalmente abordado por Turing (1950), onde o autor indica que a máquina deve ser capaz de mudar para aprender. O termo *machine learning* foi popularizado por Samuel (1959), onde é feita uma comparação entre duas metodologias de aprendizado de máquina considerando o problema do jogo de xadrez. Os dois métodos de AM são correlatos à rede neural artificial, e a máquina aprendeu a partir de uma rotina de recompensa/punição. A evolução das técnicas de aprendizado de máquina está intimamente relacionada com a capacidade de processamento de dados. Consequentemente, o desenvolvimento das técnicas de *machine learning* esteve atrelado à capacidade de processamento dos computadores. Segundo Georgiev (2021), o interesse em aprendizado de máquina teve um forte crescimento entre os anos 2010 e 2020 e no presente momento já alcançou o pico de interesse.

Existem diversos métodos de aprendizado de máquina, com variações e melhoramentos plenamente adaptáveis. Os métodos de aprendizado podem ser subdivididos em três categorias: Métodos não-supervisionados, supervisionados e por reforço. Nos métodos não-supervisionados a resposta é gerada a partir das relações internas dos próprios dados (ROUSSEEUW, 1987; SAMMUT; WEBB, 2011; ALPAYDIN, 2014). Os métodos supervisionados tendem a copiar um procedimento para gerar uma resposta, e por isso, precisam aprender com os dados de treinamento (KOUTROUMBAS; THEODORIDIS, 2008; SAMMUT; WEBB, 2011). Por fim, o aprendizado por reforço utiliza o conceito de punição ou recompensa para gerar o modelo de sistema. Usualmente, as técnicas de aprendizado por reforço são mais interessantes em recriar a interação com o sistema, e não na modelagem deste (SAMMUT; WEBB, 2011; KEESMAN, 2011).

É bem provável que o método de aprendizado mais conhecido seja a regressão linear, que utiliza o conceito de hiperplano, uma figura geométrica sem curvatura com $n-1$ dimensões em um espaço de n dimensões. A regressão linear tende a ajustar os dados neste hiperplano para gerar o modelo de sistema. Também existem os métodos que utilizam um hiperplano para subdividir o espaço a fim de executar destacar classes, como o perceptron e a máquina de vetores de suporte (SVM do inglês) (SMOLA; SCHÖLKOPF, 2004; BISHOP, 2006). Alguns métodos como o K-ésimo Vizinho mais Próximo e o *K-means* utilizam a distância Euclidiana para realizar o agrupamento (ROUSSEEUW, 1987; ARTHUR; VASSILVITSKII, 2007). A árvore de decisão e métodos associados (floresta aleatória, C4.5

[1] Não específico à arquitetura de Von Neumann

e as árvores aprimoradas) utilizam o conceito de eventos sequenciais para estruturar o aprendizado (MOLA, 1998; SAMMUT; WEBB, 2011; PEDREGOSA et al., 2011a; HALL; HALL, 2017). Alguns métodos de inferência estatística também são considerados como métodos de aprendizado de máquina, e o exemplo mais conhecido é o naïve Bayes, que emprega o teorema de Bayes na classificação (MURPHY, 2012; BRUCE; BRUCE, 2019). Por fim, o conceito de camadas pode ser aplicado a qualquer tipo de método, e seus exemplos mais comuns são o perceptron multicamadas e as redes neurais artificiais (BISHOP, 2006; SAMMUT; WEBB, 2011; PEDREGOSA et al., 2011b).

As redes sociais são as que mais investem neste tipo de tecnologia, pois o aprendizado de máquina está diretamente relacionado à maximização do lucro através da relação entre usuários e anunciantes. Neste âmbito, os métodos de AM são utilizados no agrupamento de usuários em categorias que estejam diretamente relacionadas a produtos específicos (MARR, 2016; ABDULKADER; LAKSHMIRATAN; ZHANG, 2016; MEDVEDEV; WU; GORDON, 2020). Outra área importante em que o AM atua fluentemente é a robótica. Por exemplo, veículos autônomos dependem de sensores e câmeras para monitorar o ambiente ao redor, e os métodos são utilizados para uma melhor integração entre esses sensores (ORS, 2020). Na área da exploração espacial, o principal desafio está na escassez de dados que são essenciais para a utilização dos métodos. Para facilitar o deslocamento do astromóvel Curiosity, a NASA conta com o auxílio popular, onde o público ajuda a reconhecer possíveis obstáculos no solo marciano (ONO et al., 2020).

Nas geociências, os métodos de AM têm aplicações bem variadas. Dell'Aversana, Ciurlo e Colombo (2018) utilizam várias técnicas de *machine learning* para integrar dados sísmicos, eletromagnéticos, gravimétricos e de poços na inversão e modelagem geológica. Jia et al. (2021) Usaram o estaqueamento, uma aplicação que consiste na integração de várias métodos de AM, para gerar um modelo geológico 3D utilizando dados de densidade residual, susceptibilidade magnética e da geologia considerando amostras de calha (rocha residual da perfuração). A precisão de todos os resultados foi superior a 90%. Bray e Link (2015) realizaram uma análise comparativa das técnicas de rede neural, floresta aleatória e SVM para localizar munição não detonada através do campo magnético total. Neste caso, 100% dos objetos foram detectados com uma taxa de falsos positivos de 28%. Yang et al. (2020) utilizaram o agrupamento difuso, uma técnica de aprendizado de máquina para aprimorar o resultado da modelagem Magnetotelúrica no processo da inversão iterativa.

Kuang, Yuan e Zhang (2021) criaram um método de *deep learning* para descobrir a localização do mecanismo focal, que descreve a deformação no foco durante um terremoto. O *deep learning*, ou aprendizado profundo em tradução livre, é um método de AM usualmente associado às redes neurais, com grande enfoque na estruturação e na quantidade de camadas. O teste considerou dados sintéticos e estimou com sucesso quatro terremotos com magnitudes superiores a 5.4 Mw e que pode ser utilizado em regiões com ou sem

histórico de sismos. A grande vantagem reside na celeridade do processo para predizer a origem do mecanismo focal. Wrona et al. (2018) testaram 20 técnicas de AM no auxílio à interpretação sísmica. A metodologia consistiu em usar uma classificação manual feita em um volume sísmico, e utilizar essa classificação para os métodos de AM. A melhor técnica, uma variante do SVM, foi utilizada na classificação do conjunto completo de dados. Os resultados apresentaram uma precisão de 98 a 99%. Akkaş et al. (2015) usaram uma variante da árvore de decisão denominada C5.0 como uma alternativa rápida e eficiente para identificar e classificar minerais a partir dos dados do espectrômetro de raios-x e do microscópio eletrônico de varredura. Pradhan (2013) Utilizou três algoritmos de AM árvore de decisão, SVM e um sistema de inferência *neuro-fuzzy* adaptável na identificação de possíveis regiões suscetíveis a deslizamentos de terra através de fotos aéreas. O *neuro-fuzzy* ou neuro difuso também é uma variação do método de redes neurais, com aplicação da lógica *fuzzy*. Dos testes feitos, a árvore de decisão foi o método que apresentou a maior quantidade de acertos, e a inferência *neuro-fuzzy* foi o que apresentou menos falso positivos.

Métodos de aprendizado de máquina também são amplamente utilizados em dados de poços. Neste tipo de dado, o problema consiste em associar uma amostragem de rocha com as respostas físicas de um instrumento, e produzir uma interpretação que seja condizente com a geologia regional. A amostragem, em geral, consiste de fragmentos de rocha que são subprodutos da etapa da perfuração do poço, ou amostragens pontuais após a perfuração em profundidades específicas. A resposta física é obtida em intervalos regulares de profundidade por instrumentos que captam respostas físicas das rochas, sendo estas induzidas ou não (ELLIS; SINGER, 2007). Não existem atalhos ou padrões a serem memorizados para a interpretação, e cada poço/região representa um desafio único (ASQUITH; GIBOSN, 1982). Diversos métodos de interpretação automática surgiram para sobrepor este tipo de problema. Delfiner, Peyret e Serra (1987) utilizaram a análise discriminante, para correlacionar os valores de perfis às fácies litológicas. A análise discriminante é um método estatístico, usualmente utilizado na redução do número de parâmetros, e que atua em agrupamentos. Moradi, Tokhmechi e Pedram (2017) compararam a performance de duas metodologias de aprendizado de máquina na elaboração de perfis de fácies. O treinamento foi feito a partir de descrições de amostras pontuais do poço e lâminas petrográficas. Os resultados indicaram que o método KNN (sigla em inglês para K-ésimo Vizinho mais Próximo) apresentou melhores resultados que o naïve Bayes. Li, Chan e Nguyen (2013) utilizaram técnicas de AM na predição da produção de poços considerando os valores existentes. Foram testadas as redes neurais artificiais, uma variante da árvore de decisão denominada C4.5 e uma rede neural em árvore, sendo que esta última produziu os melhores resultados. Métodos de AM também podem ser utilizados para gerar perfis sintéticos como no trabalho de Akinnikawe, Lyne e Roberts (2018), onde foram utilizadas diversas técnicas de aprendizado para estimar os perfis fotoelétricos (PE) e de resistência a pressão não confinada (UCS). Apesar de ser possível simular estes perfis em laboratório, os autores

argumentam que os perfis sintéticos são uma opção mais viável economicamente. O dado original foi dividido na relação de 70:30, onde 70% foi utilizado para o treinamento e 30% para validação onde os melhores resultados foram as técnicas de redes neurais e floresta randômica. Carreira, Ponte-Neto e Bijani (2018) aplicaram o Mapa auto-organizável e os classificadores de Mahalanobis na identificação de litofácies, seguindo uma linha supervisionada. Os testes foram executados em uma seção geológica sintética. Por fim, Hall e Hall (2017) apresentam os resultados de uma competição de AM onde o objetivo era a predição de 9 facies através do aprendizado supervisionado em dados de poços previamente interpretados, utilizando 7 perfis geofísicos. O melhor resultado foi de uma variante da árvore de decisão denominada árvores aprimoradas (*Boosted trees*).

1.1 Motivação científica

Dentre os diversos métodos mencionados até aqui destacamos o *naïve* Bayes que é uma metodologia de AM baseada no teorema de Bayes e é uma das versões mais simples dos estimadores Bayesianos, (CASELLA; BERGER, 2010; LINDBERG; RIMSTAD; OMRE, 2015; LI; ANDERSON-SPRECHER, 2006). Notoriamente, a performance do NB está diretamente relacionada aos dados a serem classificados. Portanto, consideramos que o naïve Bayes é um método adequado para a solução de problemas de classificação de eletrofácies a partir de perfis geofísicos de poços, pois este tipo de dado possui algum ruído aleatório referente aos instrumentos, assim como uma convolução[2] que ocorre no processo de aquisição (ASQUITH; GIBOSN, 1982; LINDBERG; RIMSTAD; OMRE, 2015). Sobre sua performance, trabalhos como Horrocks, Holden e Wedge (2015), Xie et al. (2018), Cracknell e Reading (2014) e Cheraghi, Kord e Mashayekhizadeh (2021) indicam que o *naïve* Bayes apresenta os piores resultados quando comparado com outras metodologias de AM. Outros trabalhos como Gonçalves et al. (2017) e Handhal et al. (2022) indicam uma performance razoável; E em Hassan, Rafi e Shaikh (2011), Ashari, Paryudi e Tjoa (2013) e Yadav e Thareja (2019) o NB desponta como o melhor método. Zhang (2004) salienta que o *naïve* Bayes é um dos métodos mais eficientes e eficazes em se tratando de aprendizado indutivo (guiado por exemplos). Portanto, entendemos que existem questionamentos ainda vigentes sobre o desempenho desta metodologia, conforme alguns citados por Li e Anderson-Sprecher (2006) como, por exemplo, a baixa performance da versão não paramétrica do NB associada ao sobreajuste (*overfitting* do inglês). Acerca do desempenho, uma análise nas referências supracitadas indica que o NB, quanto a acurácia dos resultados, é superior a metodologias como o KNN e a árvore de decisão, mas é inferior quando comparado com a floresta aleatória, e com os métodos de rede. Baseados nesta motivação, foi efetuada uma análise comparativa considerando o classificador *naïve* Bayes e quatro estratégias para aprimorar o cálculo das probabilidades a *priori* e as verosimilhanças que compõem o classificador

[2] sobreposição entre leituras; também conhecido como *shoulder effect*

NB. Essas estratégias, juntamente com o método NB, foram aplicadas no problema da classificação de eletrofácies a partir de perfis geofísicos de poços no campo de Massapê, Bacia do Recôncavo.

A primeira estratégia, aqui denominada de KDE, referente a sigla homônima para estimativa da densidade do kernel (KDE em inglês), estabelece a relação entre os dados e a litologia de modo não paramétrico, utilizando este conceito para produzir um melhor ajuste entre as verossimilhanças e a distribuição do dado de treinamento. Li e Anderson-Sprecher (2006) realizaram este primeiro comparativo, em dados de poços, e concluíram que o *naïve* Bayes, em sua estrutura paramétrica gaussiana, apresenta resultados melhores que o KDE. Este resultado é considerado surpreendente visto que as verossimilhanças estabelecidas pelo KDE são mais ajustadas ao dado. Xiang, Yu e Kang (2016) propõem uma nova versão do método *naïve* Bayes, considerando uma aplicação ponderada do KDE denominada de AW-SKDE, cujo propósito é aprimorar os resultados do *naïve* Bayes padrão. Foi utilizado o banco de dados de Newman et al. (1998), e os resultados foram comparados com outros seis métodos, incluindo o *naïve* Bayes com distribuição Gaussiana, onde os resultados foram similares ao do método base. Considerando estes questionamentos ainda vigentes acerca da utilização do KDE, esta estratégia foi incorporada com o objetivo de estender a discussão.

A segunda estratégia foi a aplicação do *Tuning* — termo que pode ser traduzido como afinação ou sintonização — e é referente ao ajuste de parâmetros associados ao desempenhos dos métodos, e sem relação direta com os dados. Agrawal (2021) utiliza a taxa de aprendizado para exemplificar o *Tuning*. A taxa de aprendizado é um hiperparâmetro usualmente utilizado em métodos de minimização do erro; se a taxa de aprendizado for muito alta, o algoritmo vai convergir para um ponto mínimo instável, se a taxa de aprendizado for muito baixa, o algoritmo vai precisar de muitas etapas até atingir a convergência. Portanto, é preciso encontrar um valor ótimo para a taxa de aprendizado, e é esta a tarefa do *tuning*. Horrocks, Holden e Wedge (2015) utilizaram o *tuning* na escolha do número de neurônios da camada oculta nas redes neurais artificiais; na margem e na escolha do parâmetro do kernel na máquina de vetores de suporte. Os hiperparâmetros são ajustados utilizando a validação cruzada. Xie et al. (2018) - utilizaram o *tuning* na máquina de vetores de suporte, redes neurais artificiais, floresta aleatória e no *gradient tree boosting*. Ao todo, 18 hiperparâmetros são aprimorados e testados e as amostras são separadas, primeiramente na razão 80-20, onde 80% é separado para teste e 20% para validação. A fração de 80% é utilizada para o *tuning* nos métodos previamente citados, onde os métodos são avaliados pela validação cruzada. Considerando os exemplos supracitados, considerou-se a *priori*, que é uma probabilidade mais flexível do método *naïve* Bayes, como um hiperparâmetro a ser aprimorado. Portanto, a proposta é utilizar o *tuning* para encontrar uma *priori* ótima, partindo de uma premissa aleatória. A convergência para essa *priori* foi feita através da maximização do valor-f (*f-score*), uma métrica que indica a

acurácia dos resultados.

A terceira estratégia foi a arquitetura, com base no trabalho de Horrocks, Holden e Wedge (2015). A arquitetura é um caso particular de uma análise integrada denominada *ensemble* (também conhecida por *committee machines* e *multiple classifier systems*) é uma estratégia onde o treinamento é realizado de modo fragmentado, e seus resultados finais são combinados (SAMMUT; WEBB, 2011). Aplicações *ensemble* podem ser homogêneas, quando utilizam um mesmo classificador com dados distintos ou heterogêneos quando utilizam vários métodos distintos [3] (ELISH; HELMY; HUSSAIN, 2013). Breiman e Breiman (1996) alegaram que a combinação classificadores instáveis como a árvore de decisão e as redes neurais, tende a reduzir os erros nos testes. Um algoritmo é dito como instável se ele apresenta uma grande diversidade de resultados devido a pequenas alterações no dado de treinamento (DIETTERICH, 2000). Considerando a classificação de litofácies, Tewari e Dwivedi (2020) comparam *ensembles* heterogêneos combinando os métodos perceptron multi-camadas, a máquina de vetores de suporte, e *gradient boosting* na classificação de dados de poços. A precisão dos resultados ficou entre 67% e 95.8% e que a performance do *ensembles* é dependente das metodologias de aprendizado escolhidas. Horrocks, Holden e Wedge (2015) comparam os métodos de aprendizado de máquina com suas versões utilizando o *ensemble* homogêneo (aqui denominado por arquitetura *committee*). Os resultados que a utilização desta estratégia promove um ganho de performance para os métodos. Portanto, ao considerar a *ensemble* no *naïve* Bayes, espera-se encontrar algum incremento na performance deste método, visto os resultados anteriormente apresentados.

A quarta — e última — é a maior novidade apresentada neste trabalho, consiste na aplicação do NB considerando a geologia da Bacia do Recôncavo, em particular, do Campo de Massapê. Neste campo, Freire et al. (2020) associam três zonas de produção em profundidade a eventos contíguos, que se estendem por todo campo e que compartilham de um mesmo tempo geológico. Portanto, esta última estratégia, aqui denominada como CRC, considera a aplicação do método NB restrito à zonas em profundidade. Para cada zona, foram definidas probabilidades a *priori*, que foram aplicadas no método NB. O propósito teórico considera que aplicações segmentadas (em profundidade no caso) tendem a produzir resultados mais precisos porque exploram homogeneidades locais associadas à gênese da rocha sedimentar.

Inicialmente, uma série de perfis geofísicos, incluindo os de litologia interpretada, também utilizados por Freire et al. (2020), foram solicitados, analisados e processados, por meio dos gráficos de dispersão (*crossplot* em inglês) e dos perfis de poços (*logplots* em inglês). Nesta etapa, o propósito é estabelecer o subconjunto ideal para compor o banco de dados de treinamento, cujos perfis de raios-gama (GR), sônico (DT) e de resistividade (ILD) são considerados. Em seguida, as quatro estratégias mencionadas, e o *naïve* Bayes

[3] Aplicações heterogêneas também podem ser denominadas simplesmente por *ensemble*

padrão, são treinados a partir do banco de dados de treinamento, composto por 12 poços do Campo de Massapê. Consideramos, neste trabalho, o *naïve* Bayes padrão como a versão paramétrica do NB que considera a distribuição normal ou Gaussiana. Por comodidade, denominamos esta aplicação como NB STD em referência a *standard* do inglês. Vale salientar que todas as estratégias foram desenvolvidas em Python. Para fundamentar a análise proposta, utilizamos os perfis GR, DT e ILD do poço 7MP-50D-BA como validação. Este trabalho também realiza uma aplicação real das estratégias, considerando o poço 5-BRSA-365-BA que não possui classificação litológica. Foram utilizadas a matriz de confusão, a revocação e o valor-f como métricas estatísticas de análise, e validação das classificações estabelecidas.

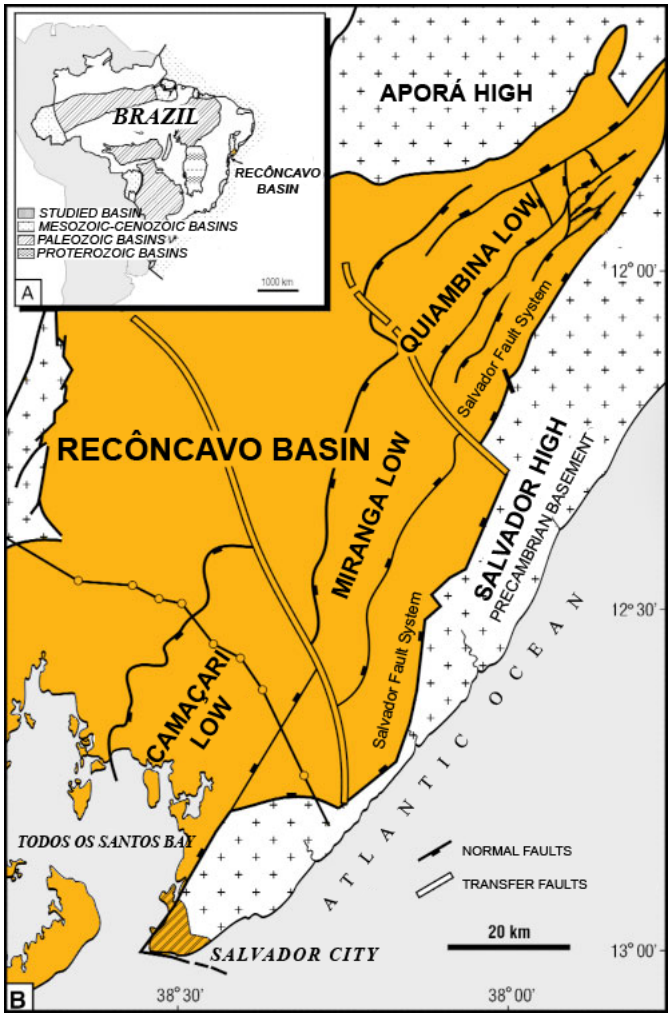
1.2 Aspectos Inovadores

Frente às possibilidades acerca do método *naïve* Bayes, foram propostas alternativas metodológicas para aprimorar os resultados obtidos pelo classificador NB. Particularmente, a estratégia para integrar informações geológicas prévias ao cálculo das probabilidades a priori é, definitivamente, uma contribuição inédita deste trabalho. Destas estratégias, duas são usualmente aplicadas a métodos de aprendizado de máquina, com o intuito de otimizar seus resultados, uma é própria para o método NB, e a última, aqui denominada de CRC, foi desenvolvida neste trabalho e considera a estratigrafia da Bacia. Todos os testes foram realizados à partir de dados de poços da Bacia do Recôncavo, uma bacia madura (em relação à exploração petrolífera), e que possui uma escassez considerável quanto à trabalhos sobre aprendizado de máquina.

2 Geologia Regional

A Bacia do Recôncavo é uma bacia sedimentar que está localizada no centro-leste do Estado da Bahia, na Região Nordeste do Brasil, e possui uma área aproximada de 11.500 km^2. Os estudos nestas Bacia são datados desde meados do século XIX e sua exploração se iniciou em 1937 com a descoberta do campo de lobato, na época sob designação do antigo Conselho Nacional do Petróleo (CNP) e cujas competências atuais pertencem a Petróleo Brasileiro S.A. (PETROBRÁS). Esta bacia está oficialmente em regime de exploração desde 1941 e atualmente é classificada como bacia madura, o que significa que seus campos já atingiram o auge de sua produção e hoje estão em fase de declínio (PRATES; FERNANDEZ, 2015; PETRÓLEO BRASILEIRO S. A., 2020).

Figura 1 – Localização da Bacia do Recôncavo em laranja e suas feições estruturais principais (BRUHN, 1999).



2.1 Estruturação e Embasamento

O Sistema de Riftes Recôncavo-Tucano-Jatobá é um aulacógeno (rifte abortado) que está relacionado a fragmentação do Paleocontinente Gondwana durante o Neojurássico – Eocretáceo, e que foi extinto no Aptiano (final do Eocretáceo). A Bacia do Recôncavo está localizada na porção sul deste sistema de riftes e é limitada a norte e noroeste com a Bacia de Tucano, pelo Alto de Aporá, a sul com a Bacia de Camamu pelo sistema de falhas da Barra; a leste pelo sistema de falhas de Salvador; e pela Falha de Maragogipe a oeste Figura 1. Esta bacia é um meio-gráben, ou seja, apresenta uma estrutura basculamento, com uma borda de falha bem pronunciada a leste (sistema de falhas de Salvador) e uma borda flexural a oeste Figura 2. O embasamento da bacia é composto por rochas arqueanas do Bloco Serrinha, a oeste e norte; pelos cinturões Itabuna-Salvador-Curaçá, a oeste-sudoeste e Salvador-Esplanada, a leste-nordeste. Ao norte é composto por rochas de idade Neoproterozoica, relacionadas ao Grupo Estância (NETTO; OLIVEIRA, 1985; CARLOTTO, 2006; COURA, 2006; GORDON; DESTRO; HEILBRON, 2017; NELIZE, 2011).

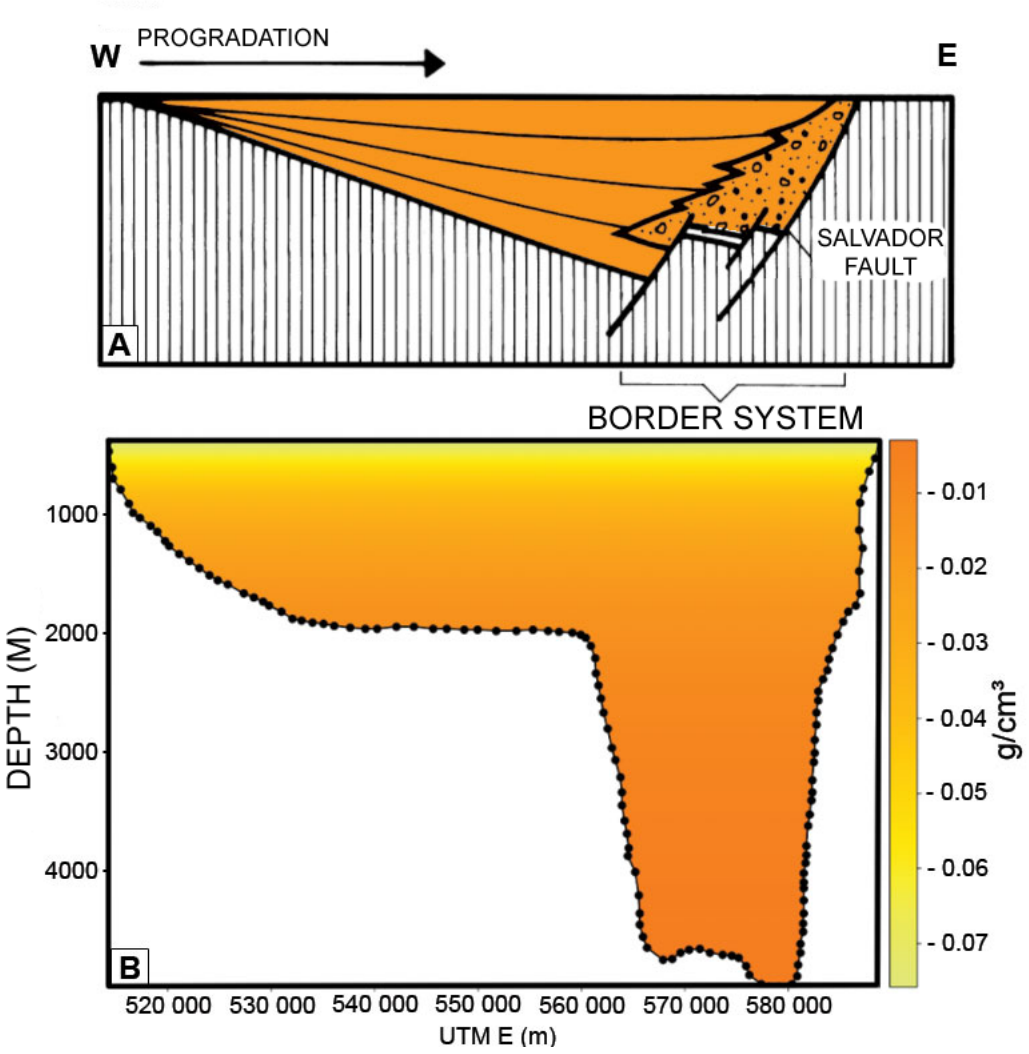
2.2 Evolução Sedimentar

A Bacia do Recôncavo possui registro sedimentar desde o Paleozoico até o recente, sendo que grande parte do preenchimento é referente ao Eo-Cretáceo (OLÍVIO et al., 2007). Uma síntese de toda a evolução da bacia está apresentada na Figura 3.

Paleozóico - Os primeiros registros sedimentares da região são datados do Paleozóico (Permiano) na Formação Afligidos e compreendem um ambiente transicional marinho raso e desértico com lagunas salinas do Membro Pedrão sobrepostos por um sistema lacustre do Membro Cazumba (CAIXETA et al., 1994; OLÍVIO et al., 2007).

Mesozóico - O registro Jurássico corresponde ao Grupo Brotas composto pelas Formações Aliança e Sergi que correspondem a uma deposição majoritariamente flúvio-eólica, com exceção ao Membro Capianga da Formação Boipeba cuja deposição é majoritariamente lacustre. A Formação Sergi também indica a presença de sistemas fluviais entrelaçados com posterior retrabalhamento eólico (CAIXETA et al., 1994). Na base do Grupo Santo Amaro a Formação Itaparica corresponde a uma nova transgressão lacustre (predomínio do ambiente lacustre) que recobre os sedimentos da Formação Sergi. Por fim, a Formação Água Grande é referente aos últimos depósitos fluviais com retrabalhamento eólico e é similar à Formação Sergi. A Formação Candeias corresponde ao preenchimento inicial da Bacia do Recôncavo, composto predominantemente por um ambiente lacustre profundo onde a Bacia do Recôncavo é caracterizada como bacia faminta, ou seja, onde a geração de espaço de acomodação excede o suprimento sedimentar (BRUHN, 1999; GORDON; DESTRO; HEILBRON, 2017). Essa sedimentação lacustre é representada pelo Membro Tauá, enquanto que o Membro Gomo é mais característico de depósitos de correntes de turbidez(NETTO;

Figura 2 – A - Seção da Bacia do Recôncavo esquemática (MAGNAVITA; SILVA, 1995); B - estimativa de profundidade da bacia por modelagem gravimétrica (SALES; BIJANI; FREIRE, 2019).

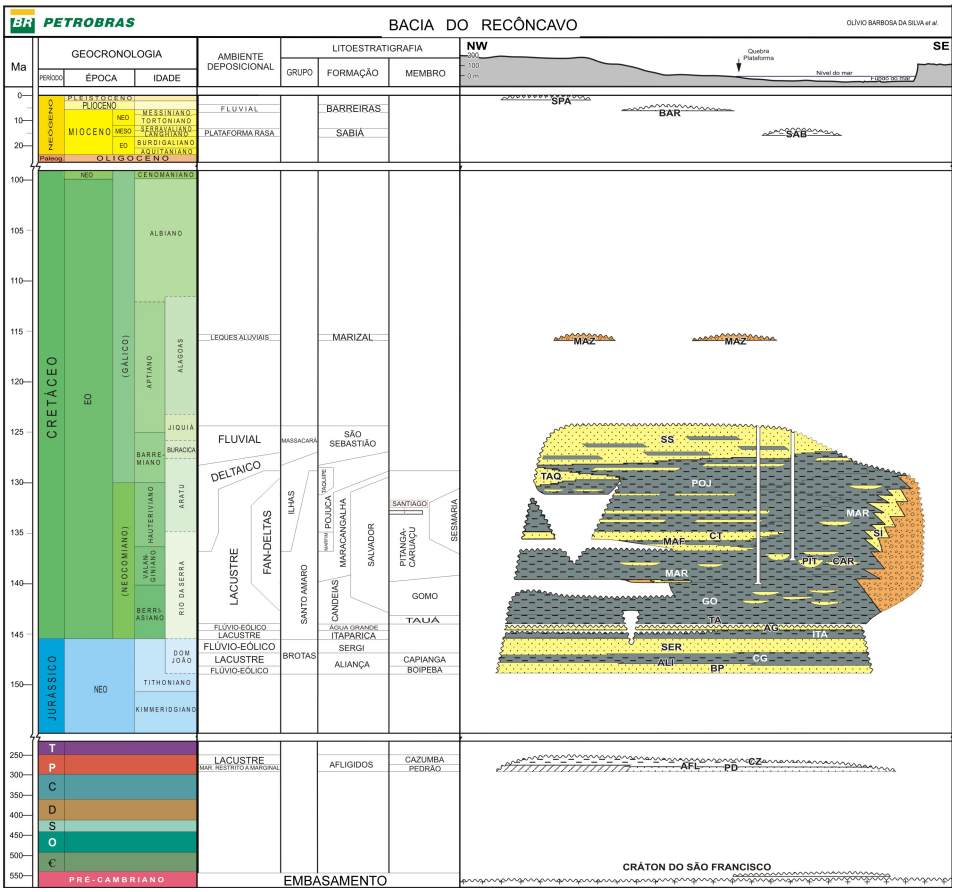


OLIVEIRA, 1985; COURA, 2006; CARLOTTO, 2006; GORDON; DESTRO; HEILBRON, 2017). A Formação Maracangalha é a última formação onde existe o predomínio do ambiente lacustre, e dentro dela encontram-se corpos de arenitos de origem turbidítica referente aos Membros Caruaçu e Pitanga. A Formação Maracangalha se distingue da Formação Candeias por ser um período de estabilidade tectônica e que indica o fim da fase de expansão da bacia, e apresenta no topo uma tendência regressiva (raseamento). Na borda sudoeste da bacia, próximo a falha de salvador, encontra-se a Formação Salvador que são depósitos grosseiros de leques aluviais e movimentos de massa referentes a borda de falha e que se estendem por grande parte do Neocomiano. O Grupo Ilhas Indica um ambiente predominantemente deltaico que migrou de norte para sul. Dentro deste Grupo a Formação Marfim caracteriza o processo de assoreamento da bacia em clima seco (CAIXETA et al., 1994; OLÍVIO et al., 2007). A Formação Pojuca também se caracteriza por uma sedimentação predominantemente deltaica, mas apresenta também registros de afogamentos do sistema

deltaico que resultam em importantes marcos estratigráficos. A última formação do Grupo Ilhas é a Formação Taquipe que corresponde a depósitos de fluxos gravitacionais associado ao Cânion do Taquipe, bem definido na região. O Grupo Massacará, bem representado pela Formação São Sebastião é formado por sedimentação predominantemente fluvial e indica a fase final do assoreamento do rifte. A Formação Marizal corresponde a depósitos de leques aluviais durante o Neo-Aptiano característico de sistemas aluviais.

Cenozóico - Correspondentes aos depósitos mais recentes, a Formação Sabiá corresponde a um afogamento miocênico regional e a Formação Barreiras indica um ambiente predominantemente aluvial.

Figura 3 – Carta estratigráfica adaptada da Bacia do Recôncavo segundo Olívio et al. (2007).



2.3 Formação Maracangalha e Membros Caruaçu e Pitangas

Os depósitos lacustres da Formação Maracangalha são formados principalmente por folhelhos cinza-esverdeados a cinza-escuros, característicos de lago profundo e com algumas ocorrências de margas (S; L, 2003; SANTOS; CORREA-GOMES, 2018). Já os membros Caruaçu e Pitanga são depósitos resultantes de processos gravitacionais, fluxos turbidíticos e fluxos hiperpicnais, de modo que o Membro Caruaçu é caracterizado por ser um reservatório

petrolífero de melhor qualidade que o Membro Pitanga. Existe uma grande variedade de fácies associadas a estes tipos de depósitos e várias interpretações (COURA, 2006), mas neste trabalho, optamos por considerar as quatro eletrofácies designadas por Freire et al. (2020). Eletrofácies é um termo designado para caracterização de fácies a partir de propriedades físicas das rochas (em particular para os perfis de poços). As eletrofácies possuem uma resolução bem menor e são mais limitadas do que qualquer análise de rocha em situ ou em laboratório, e portanto, só permitem uma interpretação limitada. As eletrofácies descritas por Freire et al. (2020) e suas descrições são citadas a seguir:

1. Folhelho;

   correspondente a sedimentação fina, de baixa energia, formada em ambiente redutor e originada majoritariamente pela deposição lacustre profunda. Esta eletrofácies corresponde a Formação Maracangalha que se encontra intercalada com as fácies dos membros Caruaçu e Pitangas (COURA, 2006; FREIRE et al., 2020).

2. Siltito;

   Uma eletrofácies que representa uma rocha composta pelo mineral silte (em maior parte), que corresponde a uma energia de deposição superior a das argilas e inferior aos grãos de areia. Esta rocha é resultante da decantação de um fluxo quase estático característico da porção final de fluxos turbulentos (COURA, 2006; SANTOS; CORREA-GOMES, 2018).

3. *Slurry* (ou Arenito lamoso);

   Eletrofácies correspondente a arenitos muito finos de seleção pobre e com bandamentos internos de granulometria argilosa e que estão associados a depósitos do tipo *slurr* (escorregamento). Esta seria a principal versão, mas essencialmente o slurrie pode ser qualquer arenito fino, moderado a mal selecionado e lamoso. Segundo Freire et al. (2020) a principal característica do *slurry* é sua porosidade que varia de 5% a 9% (S; L, 2003; COURA, 2006; SANTOS; CORREA-GOMES, 2018; FREIRE et al., 2020).

4. Arenitos

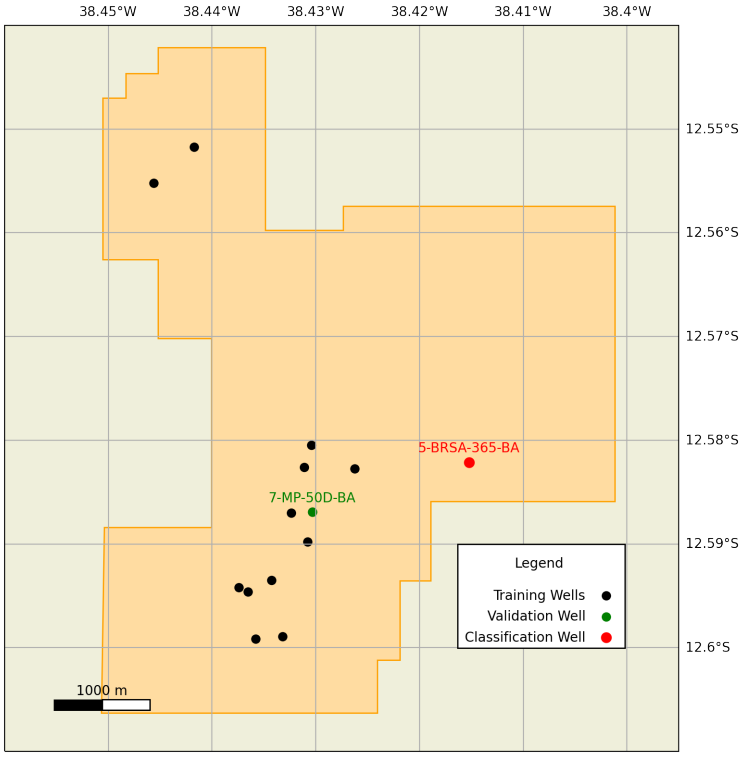
   As rochas que compõem a eletrofácies Arenito seriam os arenitos finos a medios com seleção moderada ou moderada a pobre, mas com porosidades superiores a 9%(COURA, 2006; FREIRE et al., 2020).

Também foram consideradas três zonas de produção denominadas de CR-1, CR-2 e CR-3 cuja gênese está relacionada à sistemas turbidíticos segundo Mutti e Normark (1991). Estas zonas são identificadas em relação à razão arenito/folhelho, iniciando em cerca de 0.5 para o CR-1, e aumentando gradativamente até o CR-3.

2.4 Perfis Geofísicos de Poços

Os dados desse trabalho foram cedidos pela Petróleo Brasileiro S.A. (PETROBRAS), por meio do projeto de pesquisa e desenvolvimento financiado, em parceria com a Universidade Federal Fluminense, e são são oriundos do Campo de Massapê um campo petrolífero com área aproximada de 23,97 km^2 e que está localizado próximo ao Município de São Sebastião do Passé (BA). Neste campo, o principal reservatório está localizado no Membro Caruaçu, mas também existem reservas importantes nas Formações Pojuca e Marfim (PRATES; FERNANDEZ, 2015). Deste campo foram utilizados dados de 14 poços (12 para treinamento, um para validação e um para classificação) conforme pode ser visto na Figura 4.

Figura 4 – Localização dos poços no Campo de Massapê



3 Metodologia

O fluxo de trabalho aplicado consiste de quatro etapas: processamento, para separar e estrurar os perfis GR, DT e ILD; aplicação das estratégias associadas ao classificador *naïve* Bayes em todos os poços, com destaque para o de validação (7MP-50D-BA) que foi escolhido convenientemente para salientar o desempenho da estratégia inovadora apresentada neste trabalho; a validação destes métodos com base em análises de erros; e a classificação em um poço sem litologia interpretada. A descrição detalhada de cada etapa é apresentada nos tópicos a seguir:

1. Processamento

   Corresponde a etapa de gerenciamento dos dados para a supervisão adequada dos métodos considerados neste trabalho. A sequência adotada consiste em separar os perfis desejados (DEPTH, CALI, GR, ILD, DT, e LITO) para todos os 12 poços disponíveis e separar os intervalos de profundidade referentes ao reservatório para cada um deles. Essa informação sobre os intervalos de profundidade foi proveniente dos trabalhos de Silva et al. (2018) e Da Silva et al. (2019). Em seguida, intervalos de profundidade com ausência de dados foram desconsiderados. Na sequência foram removidos todos os dados onde o perfil CALI apresentava uma variação superior a uma polegada. Nessas condições o contato entre os instrumentos e a parede do poço é reduzido, o que gera informações imprecisas (ASQUITH; GIBOSN, 1982; ELLIS; SINGER, 2007). Uma correção também foi aplicada no perfil DT com o objetivo de remover o efeito da compactação com o incremento da profundidade (ELLIS; SINGER, 2007). Após as correções e delimitações em profundidade, os dados de treinamento foram verificados manualmente, com o objetivo de encontrar os poços que possuíam uma maior e menor proporção de arenitos (rocha reservatório) e definir aquele que seria mais representativo no intervalo do reservatório para a etapa de validação. Os poços com a maior e menor quantidade de reservatório foram definidos como cenários probabilísticos, e suas proporções de eletrofácies foram calculadas para uso posterior no *naïve* Bayes. O banco de dados de treinamento foi construído, portanto, com todos os poços, excluindo o 7MP-50D-BA separado para a etapa de validação.

2. Aplicação das Estratégias

   Nesta etapa o método *naïve* Bayes é aplicado no conjunto de dados, assim como as estratégias associadas a este método. A sequência se resume, primeiramente, as análises dos gráficos de dispersão, que são um indicativo da viabilidade do método considerando os dados utilizados; posteriormente são definidos os parâmetros referentes a etapa de treinamento do NB e de cada estratégia.

3. Validação

   A penúltima etapa, de validação, consiste em comparar a classificação obtida por cada um dos métodos com a litologia interpretada. Este processo é realizado para todos os poços do Campo de Massapê [1]. Todos os resultados são agrupados em duas tabelas que indicam (para cada poço) a acurácia das aplicações. Sequencialmente, um poço representativo foi escolhido com o intuito de avaliar, em detalhe, a capacidade das estratégias na classificação de rochas.

4. Classificação

   Por fim, na última etapa, o método *naïve* Bayes é aplicado juntamente às estratégias na classificação do poço 5-BRSA-365-BA. Este poço não possui litologia interpretada visto que o perfil NPHI, utilizado no processo de classificação, está ausente.

3.1 Dados de Poços

Durante a perfuração de um poço (perfilagem) vários instrumentos acoplados à broca fazem medições de propriedades físicas do poço. Os dados foram disponibilizado no formato LAS, que é um arquivo estruturado como um cabeçalho e uma matriz, onde o cabeçalho contém informações particulares sobre o processo de perfuração, dados geográficos, e sobre as instrumentação utilizada na medição de propriedades físicas dos poços; a matriz é o dado propriamente dito, onde o número de colunas é o número de propriedades aplicadas na amostragem, considerando também uma coluna de profundidade como referência, e numero de linhas é referente ao número de amostragens realizadas em profundidade pelas ferramentas Struyk e Karst (2014). Os dados foram originalmente pré processados considerando apenas a profundidade referente ao reservatório do Membro Caruaçu.

Os objetivos dessas medições são: Garantir a segurança estrutural do poço, orientar a perfuração e calcular propriedades do reservatório que indicam o volume e a vazão de hidrocarbonetos. Estas propriedades físicas são medidas em intervalos regulares de profundidade e são denominadas de perfis geofísicos. A seguir apresentamos os principais perfis geofísicos utilizados neste trabalho (ELLIS; SINGER, 2007; ASQUITH; GIBOSN, 1982).

1. Caliper (CAL);

   Perfil que mede o diâmetro do poço em polegadas (*in*) e que é utilizado para informar propriedades estruturais do poço e a qualidade dos outros perfis.

2. Raios Gama (GR);

[1] O poço a ser validado sempre é removido (momentâneamente) do banco de dados na etapa de treinamento.

Perfil que mede a radioatividade natural emitida pelos isótopos radioativos de urânio, tório e potássio presentes nos minerais, e sua unidade de medida é graus API ($gAPI$), uma unidade padronizada pelo *American Petroleum Institute* (API). Este perfil é utilizado para quantificar o volume de argila nas rochas e também para indicar a litologia no intervalo.

3. Indução Profunda (ILD)

   O perfil de Indução, também chamado de resistividade, mede a resistividade das rochas na unidade $ohm.m$. Este perfil é usualmente utilizado na identificação dos fluídos nos poros das rochas. Os valores do perfil ILD podem variar de 0 a mais de 10.000 $ohm.m$, e portanto, para torná-lo mais similar a escala dos outros perfis, utilizamos o logaritmo do ILD (ou log(ILD)).

4. Densidade (RHOB ou ρ_b)

   Verifica o valor da densidade nas rochas usualmente na unidade g/cm^3 e é bem utilizado na determinação da porosidade. Em conjunto com outros perfis, pode ser utilizado também na identificação de litologias.

5. Neutrão (NPHI ou ϕ_N)

   O neutrão, ou porosidade neutrônica, é um perfil de porosidade que é bem sensível a presença de hidrocarbonetos nos poros das rochas. Usualmente sua unidade é adimensional, mas também pode ser apresentado em porcentagem. Além da porosidade, este perfil também pode ser utilizado em conjunto com outros para identificar litologias ou localizar hidrocarbonetos.

6. Sônico (DT)

   O perfil sônico mede o tempo de trânsito na rocha, e possui unidade $\mu.s/ft$ (microssegundo por pé). Este perfil costuma ser utilizado no cálculo das porosidade, na detecção de fraturas, mas sua principal atribuição é na calibração da sísmica. Neste trabalho, o perfil DT apresentou uma forte tendência em aumentar conforme a profundidade. Esta tendência foi removida ao se fazer a diferença deste perfil com uma linha calculada que indicava esse incremento com a profundidade. A diferença entre o perfil DT original, pela linha do efeito de compactação, somada a um valor médio, resultou no perfil DT corrigido, utilizado neste trabalho.

7. Eletrofácies

   A litologia interpretada foi feita seguindo a descrição de Dos Santos (2019), Freire et al. (2020), que corresponde aos procedimentos: a. - cálculo da curva DRDN

na i-ésima profundidade pela equação:

$$DRDN_i = \left(\frac{\rho_{b,i} - 2}{0.05}\right) - \left(\frac{0.45 - \phi_{N,i}}{0.03}\right). \tag{3.1}$$

b. - as quatro fácies do Membro Caruaçu (arenito, folhelho, siltito e *slurry*) possuem relação linear com a curva vsh, pois, quanto menor a granulometria da rocha, maior o valor neste perfil. Outro fator a ser considerado neste perfil é que os valores positivos são correspondentes aos folhelhos e siltitos, enquanto que os negativos são correspondentes ao arenito e ao *slurry*. É imprescindível salientar que as quatro fácies foram originalmente calculadas e, posteriormente, foi feita uma junção entre *slurry* e siltito, visto que os perfis utilizados como parâmetros (GR, DT e log(ILD)) não possuem resolução para discriminar estas eletrofácies.

8. Zonas

   Neste trabalho consideramos também as zonas de produção denominadas CRC-1, CRC-2 e CRC-3 que são as mesmas definidas em Mutti e Normark (1991), Freire et al. (2020), só que estendidas de modo que uma zona termine na outra. Essa extensão garante um maior intervalo de atuação da estratégia que considera o zoneamento.

3.2 Banco de Dados de Treinamento

O banco de dados foi construído a partir dos dados de poços apresentados no capítulo 2 após o processamento apresentado no fluxo de trabalho (Figura 5).

Figura 5 – Banco de dados usado considerando os 12 poços de validação do Campo de Massapê. A litologia está apresentada na forma numérica onde 57 é o código pro folhelho; 49 para o arenito; e 25,0 para o *slurry*. As zonas também estão em codificação numérica onde: 111 é referente a CRC 1; 222 é referente a CRC 2; e 333 é referente a CRC 3.

ÍNDICE	GR	DT	log(ILD)	LITO	CRC
0	64.37	58.51	1.80	57	111
1	65.84	61.21	1.73	57	111
2	66.88	64.21	1.61	57	111
3	70.35	66.76	1.48	57	111
4	75.47	68.83	1.35	57	111
5	77.60	70.48	1.25	57	111
6	78.93	71.94	1.17	57	111
7	78.03	72.93	1.10	57	111
8	76.15	73.29	01.07	57	111
...	...	...	...	...	...
201057	83.75	76.37	2.85	49	333

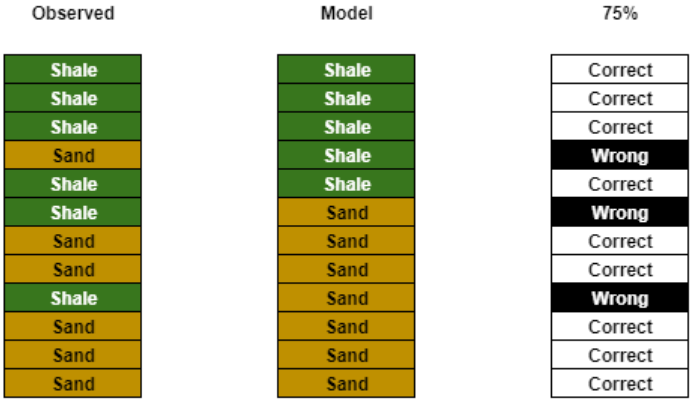
3.3 Métricas de Validação

Os modelos são usualmente avaliados em função da taxa, ou proporção de acertos, mas existem situações onde essa métrica não é suficiente. Um problema muito comum é denominado de classe rara (BRUCE; BRUCE, 2019), onde uma das classes possui uma ocorrência muito baixa nos dados observados. Como consequência, existe uma piora nos resultados gerais em decorrência da dificuldade na predição destas classes raras. Uma alternativa é utilizar outras métricas de avaliação como, por exemplo, o perfil de erros, a Matriz de Confusão e as medidas de Precisão, Revocação e Especificidade (SAMMUT; WEBB, 2011; BRUCE; BRUCE, 2019).

3.3.1 Perfil Erros

A proporção de acertos é definida como a razão do número de acertos pelo total de amostras. A Figura 6 mostra um exemplo do estruturação do perfil de erros. Vale salientar que este tipo de métrica de validação só pode ser utilizada diante do conhecimento do perfil litológico verdadeiro, o que se espera de uma aplicação com viés de verificação, no modo teste controlado.

Figura 6 – Exemplo esquemático do cálculo do perfil de erros. O primeiro perfil (à esquerda) na figura é o perfil litológico verdadeiro; o segundo (centro) consiste das litofácies classificadas; o terceiro (à direita) corresponde aos erros entre o verdadeiro e o modelado.



3.3.2 Tipos de Erros

Os erros da classificação também podem ser: verdadeiro positivo, verdadeiro negativo, falso positivo e falso negativo. Esse tipo de classificação permite um melhor entendimento sobre o desempenho do método de ML utilizado, além de permitir novas métricas de avaliação. Na Figura 7 apresentamos um exemplo esquemático que mostra como os erros são classificados considerando um conjunto de classes. É importante ressaltar que essa classificação considera sempre um problema binário e, portanto, uma classe é

sempre tomada por referência (no exemplo da Figura 7 o arenito e depois o folhelho). No caso de classes com mais de duas variáveis, como por exemplo nas três facies abordadas neste trabalho, se a referência for o arenito, o *slurry* e o folhelho correspondem a não arenito.

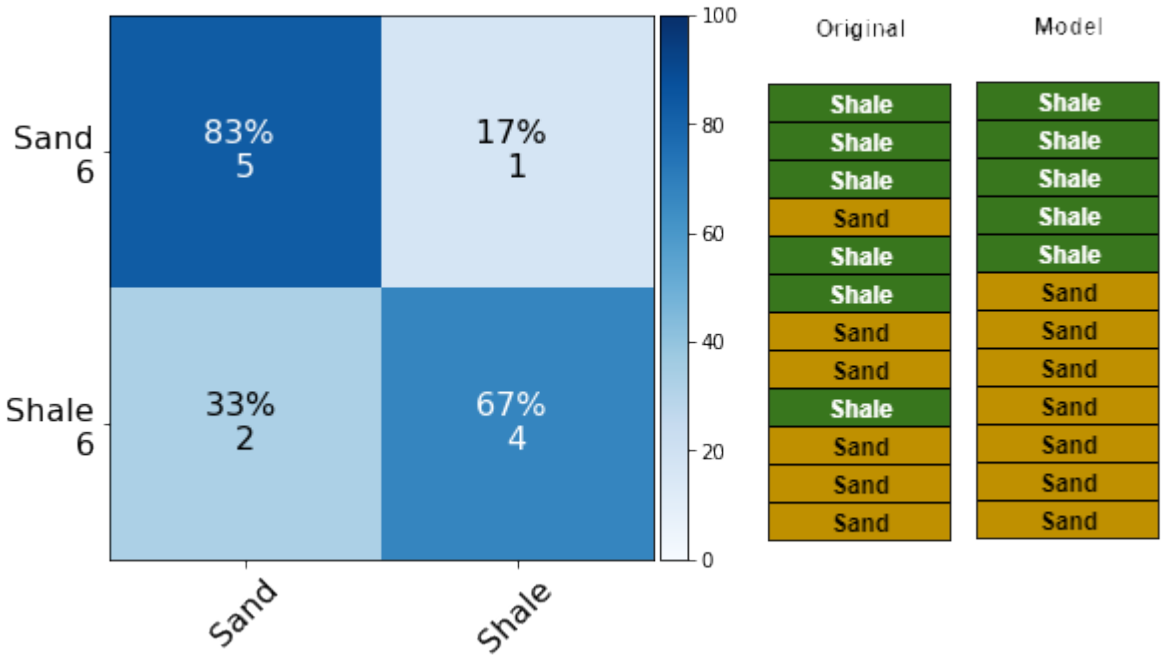
Figura 7 – Exemplo esquemático que apresenta os quatro tipos de erros: *True Negative* indica que a negativa em relação a um tipo de classificação foi correta. *False negative* indica erro na negativa em relação a classe vigente. *True positive* ocorre quando há acerto na classificação da classe e *False positive* está relacionado à afirmação em relação a uma classe for incorreta.

3.3.3 Matriz de Confusão

Nem sempre a proporção de acertos evidencia o melhor resultado, e portanto é indicado o uso de outras métricas de avaliação como a Matriz de Confusão. A Matriz de Confusão é utilizada em metodologias de classificação, ou seja, avalia a predição de classes, e consequentemente, indica a qualidade da classificação obtida. (BRUCE; BRUCE, 2019). No exemplo da Figura 8, a matriz de confusão mostra na legenda da esquerda quais são as classes presentes e sua quantidade original, ou seja, o dado observado possui seis

arenitos e seis folhelhos. A diagonal principal mostra os verdadeiros positivos, ou seja, onde era arenito e o método acertou o arenito (83%) e onde era folhelho e o modelo acertou o folhelho (67%). Na primeira linha e na segunda coluna, o valor de 17% representa o arenito que foi confundida com folhelhos, ou seja, os falsos negativos em relação ao arenito, ou falsos positivos em relação ao folhelho. Portanto, a segunda linha e primeira coluna com valor de 33% é o falso negativo em relação ao folhelho e falso positivo em relação ao arenito.

Figura 8 – Exemplo ilustrativo da matriz de confusão referente às duas litologias principais do campo de Massapê.



3.3.4 Precisão Revocação e valor F

Um conjunto de métricas muito utilizadas, em especial em ciências biológicas, são a precisão, revocação e valor-F (*Precision, Recall and Fscore* em inglês) (BRUCE; BRUCE, 2019). A precisão (PR) mede o quão exato é um resultado positivo obtido pelo modelo, e é definido por

$$PR = \frac{\sum TP}{\sum TP + \sum FP}, \tag{3.2}$$

em que TP são todos os verdadeiros positivos e FP são os falsos positivos. A revocação (RC) mede o quão sensível é o resultado em relação às classes

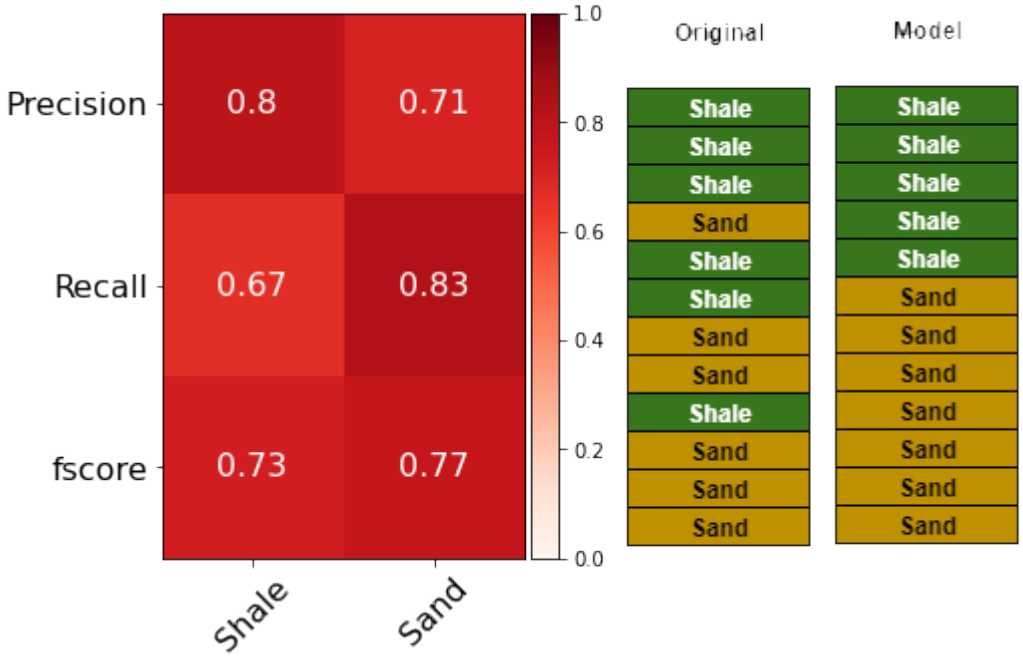
$$RC = \frac{\sum TP}{\sum TP + \sum FN}, \tag{3.3}$$

onde FN são todos os falsos negativos. A última métrica, o valor-F (VF) é uma média harmônica entre a precisão e a revocação, e sua equação é:

$$VF = 2\frac{PR.RC}{PR + RC}. \tag{3.4}$$

Na Figura 9 apresentamos um exemplo com os valores de precisão, revocação e valor-F calculados para um perfil esquemático de fácies. A primeira métrica, a precisão, indica que o modelo é mais preciso ao calcular o folhelho do que o arenito (menor número de falsos positivos), mas que o modelo é mais sensível à classe arenito (menor número de falsos negativos). Por fim o valor-F indica que o modelo gerado tem maior capacidade para prever o arenito do que o folhelho .

Figura 9 – Exemplo esquemático da *Precision* (Precisão), *Recall* (Revocação) e *fscore* (valor-f).



Neste trabalho, também utilizamos o f-score médio (ou MVF), uma métrica que permite uma avaliação mais abrangente do poço e que é definida pela seguinte equação:

$$MVF = \frac{1}{n}\sum VF_l, \tag{3.5}$$

onde n é o número de litologias e l é uma litologia qualquer. Portanto, considerando o exemplo da Figura 9, $MVF = 0.75$.

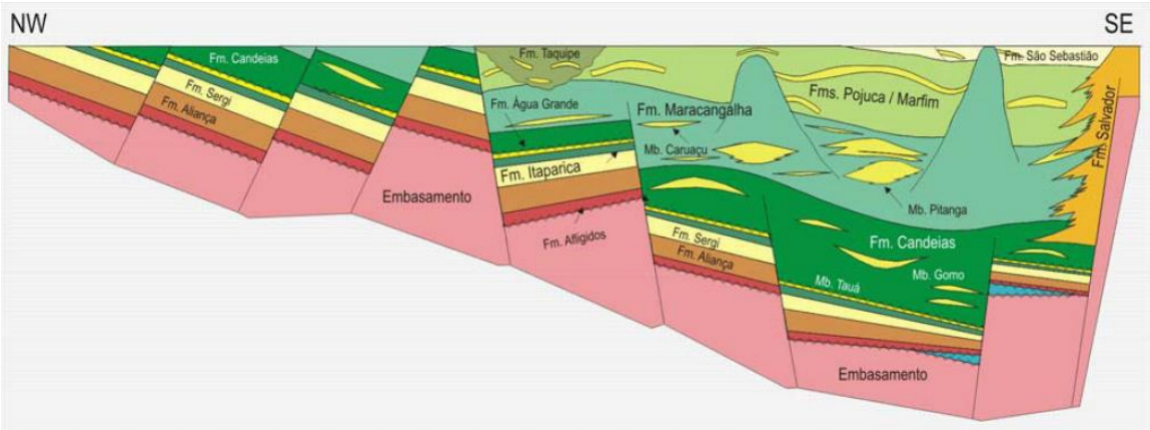
4 Teorema de Bayes

O teorema de Bayes, ou regra de Bayes, é uma formulação probabilística que justifica uma predição baseada em observações prévias e permite uma revisão (ou atualização) de uma hipótese em função de novas evidências (JOYCE, 2003; NUALART, 2011; PUGA; KRZYWINSKI; ALTMAN, 2015). Estas duas premissas permitem que a regra de Bayes venha a ser amplamente utilizada nos classificadores estocásticos (VRUGT; A., 2015; CASELLA; BERGER, 2010).

4.1 Priori

Considere o exemplo em que se deseja classificar uma amostra de rocha qualquer da Bacia do Recôncavo, e que as informações geológicas e a localização da bacia sejam as únicas informações disponíveis. Como premissa geológica em que a Bacia do Recôncavo seja uma bacia cuja origem dos sedimentos seja predominantemente terrígena (conforme descrito no Capítulo 2) um "palpite" intuitivo seria que uma amostra de rocha tomada aleatoriamente seja um folhelho, um arenito ou um *slurry*. Segundo a seção esquemática da bacia apresentada na Figura 10, o folhelho representa a maior parte das rochas desta bacia, portanto a probabilidade de que a rocha em questão seja um folhelho é a maior dentre as demais possibilidades.

Figura 10 – Seção da Bacia do Recôncavo, onde a maior parte dos sedimentos é folhelho (verde), mas também com alguma quantidade considerável de arenito (amarelo) (Da Silva, 2013).



A probabilidade atribuída ao folhelho é então considerada como um conhecimento a *priori*, isto é, uma premissa probabilística baseada em uma análise prévia sobre a geologia regional. Este pensamento pode ser formalizado se considerarmos que L seja uma variável aleatória com espaço amostral S de forma que $S = \{arenito,\ folhelho,\ slurry\}$. Portanto,

a probabilidade de uma litologia qualquer ser um folhelho, é dada por $P(L = folhelho)$, e podemos adiantar que, segundo a Figura 10, $P(L = folhelho) > P(L = arenito) > P(L = slurry)$.

4.2 Verossimilhança

Considere agora a existência de uma propriedade geofísica G aleatória, cujo espaço amostral S seja o conjunto dos números reais ($S = \mathbb{R}$), e que esta variável aleatória contínua G tende a apresentar distribuições distintas (não uniforme) considerando a litologia em L. Assim sendo, G é dependente de L, logo existe um $P(G|L)$ tal que $P(G|L) \neq P(G)$ (condição de dependência simplificada adaptada de Casella e Berger (2010)). Consequentemente, podemos inferir que $P(G|L)$ seja também um conhecimento prévio associado a cada litologia L.

4.3 Regra de Bayes

Seja a probabilidade de encontrar uma litologia $L = l$ (onde l pode ser qualquer litologia do espaço amostral de L), considerando a propriedade Geofísica aleatória $G = r$ em uma dada profundidade. Formalmente, queremos descobrir $P(L = l|G = r)$, que também pode ser descrita como probabilidade a *posteriori*. Pela regra de Bayes:

$$P(L|G) = \frac{P(G|L)P(L)}{P(G)}, \tag{4.1}$$

em que $P(G)$ é a probabilidade marginal e será discutida adiante. Momentaneamente, pode-se interpretar que a rocha não seria mais um folhelho, pois $P(L = l|G = r)$ pode ser mais baixa para o folhelho do que para as demais litologias do espaço amostral de L, o que influencia em $P(L = l)$. No entanto, se o espaço amostral de G for $\mathbb{R}$, logo, $P(G = r) = 0$, assim como $P(G = r|L = l) = 0$ onde r é um número real qualquer (MURTY; DEVI, 2011; CASELLA; BERGER, 2010). A probabilidade marginal $P(G)$ é sempre constante para qualquer valor L e pode ser redefinida conforme a lei da probabilidade total (LANGLEY; IBA, 1992; CLAPHAM; NICHOLSON, 2009; CASELLA; BERGER, 2010; UPTON; COOK, 2014)

$$P(G) = \sum_{i=1}^{n} P(G|L = l_i)P(L = l_i), \tag{4.2}$$

em que, considerando este trabalho, $n = 3$ e $l_1 = arenito$, $l_2 = folhelho$ e $l_3 = slurry$. A Equação 4.2 garante que a soma dos resultados $P(L = arenito|G = r)$ com $P(L = folhelho \mid G = r)$ e com $P(L = arenito \mid G = slurry)$ seja igual a um, garantindo a condição do segundo Axioma de Kolmogorov ($P(S) = 1$)(CASELLA;

BERGER, 2010). Portanto, resta apenas uma condição a ser resolvida que é definir o valor de $P(G = r|L = l)$.

4.4 FDP - Função de Densidade de Probabilidade

A função densidade de probabilidade (FDP) é uma função não negativa e integrável que permite transformar uma variável aleatória contínua, como G, em uma probabilidade (MURTY; DEVI, 2011; LINDBERG; RIMSTAD; OMRE, 2015). Se para uma propriedade G, observamos que é mais provável encontrar os valores r em um intervalo (a, b) do que fora deste, logo, $P(G)$ pode ser descrito por uma densidade de probabilidade. A FDP usualmente utilizada é a Gaussiana (ou normal), que é descrita pela seguinte expressão:

$$N(r) = \frac{1}{\sigma\sqrt{2\pi}}^{-\frac{1}{2}\frac{(r-\mu)^2}{\sigma^2}}, \tag{4.3}$$

onde σ é o desvio padrão, μ a média. Dentre as funções de densidade, a Gaussiana foi escolhida para esse trabalho pois é a FDP comumente utilizada em outros trabalhos (LI; ANDERSON-SPRECHER, 2006; LI; CHAN; NGUYEN, 2013) e devido ao Teorema do Limite Central onde a distribuição normal pode ser utilizada para aproximar diversas distribuições com grande amostragem (CLAPHAM; NICHOLSON, 2009; UPTON; COOK, 2014).

4.5 Condição *naïve*

Até o presente momento, apresentamos a variável aleatória G como uma propriedade geofísica de referência. Para o presente trabalho, associamos os perfis geofísicos GR, DT e log(ILD) à três variáveis aleatórias GR, DT e $log(ILD)$ (referente aos perfis homônimos), onde assume-se, teoricamente, que estas variáveis gerem valores aleatórios. Desta forma, o teorema de Bayes consiste em obter a *posteriori*:

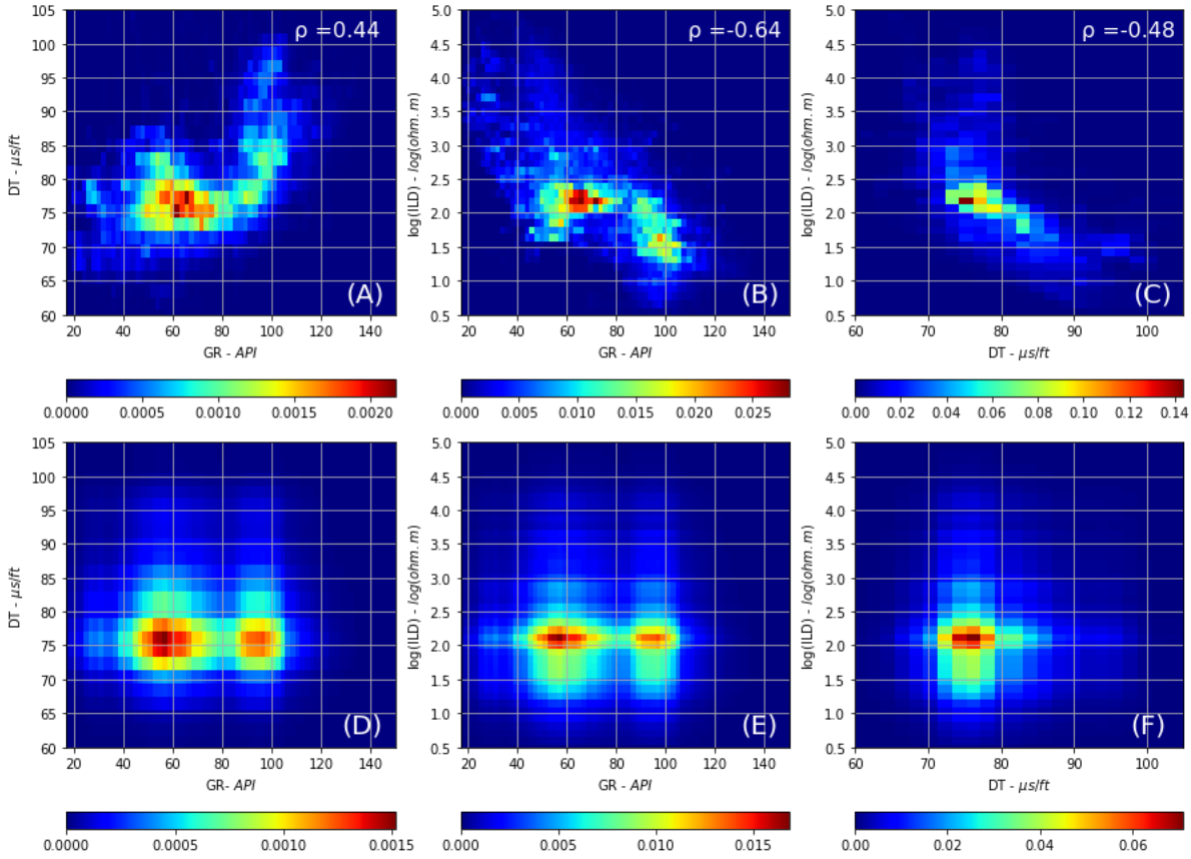
$$P(L|GR, DT, log(ILD)) = \frac{P(L)P(GR, DT, log(ILD)|L)}{P(GR, DT, log(ILD))}, \tag{4.4}$$

onde $(L|GR, DT, log(ILD))$ é um evento L condicionado à $GR \cap DT \cap log(ILD)$, considerando $\cap$ como a intersecção entre as variáveis. A condição *naïve* assume que todas as propriedades são independentes entre si, ou seja, que $GR \neq f(DT, log(ILD))$ (assim como para DT e $log(ILD)$). Em relação às probabilidades, essa independência assume que $P(GR \cap DT \cap log(ILD)) = P(GR) \times P(DT) \times P(log(ILD))$, e consequentemente (CASELLA; BERGER, 2010),

$$P(GR, DT, log(ILD)|L) = P(GR|L) \times P(DT|L) \times P(log(ILD)|L). \tag{4.5}$$

Hand e Yu (2001) apresenta uma compilação de vários trabalhos que comparam a condição *naïve* com outras metodologias, de modo que, a condição *naïve* está sempre relacionada aos melhores resultados. Hand e Yu (2001) também justifica que estes resultados favoráveis estão associados a menor variância na estimativa de probabilidades. Para auxiliar no entendimento da variância, considere o exemplo onde se deseja obter $P(GR, DT)$; segundo a condição *naïve* $P(GR, DT) = P(GR) \times P(DT)$. Estas probabilidades podem ser obtidas através de dois histogramas 1D, com n elementos cada, um para cada propriedade geofísica. Sem a condição *naïve*, $P(GR, DT)$ é um valor específico em um histograma 2D que possui n^2 elementos. Observe que, sem a condição de independência, o resultado é muito mais específico, e essa especificidade é o que aumenta a variância dos resultados. A Figura 11 (A) representa o histograma 2D para as probabilidades de $P(GR, DT)$ considerando o exemplo supracitado aplicado ao banco de dados. Já a Figura 11 (D) é o resultado de $P(GR, DT)$ considerando a condição *naïve*, para o mesmo banco de dados.

Figura 11 – Probabildiades baseadas em histogramas 2D (A,B e C) onde ρ é o coeficiente de correlação; e probabilidades segundo condição *naïve* (D, E e F).



Nas Figuras 11 (D) e (E) observam-se duas zonas bem distintas, com probabilidades acentuadas, o que não é claramente observado nas Figuras 11 (A) e (B). Estas zonas estão associadas a sistemas deposicionais distintos (turbidítico e lacustre), e demonstram a capacidade da condição de independência em salientar aspectos particulares do dado.

Outro argumento que salienta a aplicação da condição *naïve* é a inexistência da dependência real nos dados. Considerando que o grau máximo de dependência seja uma relação de uma propriedade por ela mesma (*crossplot* do GR pelo GR por exemplo), a métrica que melhor indica a dependência é o coeficiente de correlação. A Figura 11 apresenta os coeficientes de correlação ρ para as três propriedades geofísicas analisadas, comparadas duas a duas. Considerando que a independência máxima ocorre quando $\rho = 0$ e a completa dependência quando $\rho = 1$, os valores de ρ apresentados indicam uma correlação incipiente entre as propriedades, à exceção para o par GR e $log(ILD)$. É importante salientar que para a completa dependência $\rho = 1$ para relações lineares, caso contrário, $\rho \approx 1$ (ISAAKS; SRIVASTAVA", 1989; DOWDY; WEARDEN; CHILKO, 2004).

4.6 *Naïve* Bayes padrão (STD)

O classificador *naïve* Bayes é uma técnica de AM que pode ser utilizada de forma supervisionada e não supervisionada (GÁMEZ; RUMÍ; SALMERÓN, 2006). Quando utilizado de forma supervisionada, o classificador consiste na aplicação do teorema de Bayes sob a condição *naïve*.

Nesta, utiliza-se a FPD Gaussiana como padrão devido a sua capacidade de representar, uma grande variedade de distribuições (em referência ao teorema do limite central). A equação do método *naïve* Bayes para classificação é dada por (MURTY; DEVI, 2011):

$$L' = máximo \left[P(L = \bar{L}_l) \prod_{k=1}^{m} P(G = \bar{G}_k | L = \bar{L}_l) \right]_{l=1}^{n}, \tag{4.6}$$

onde L' é a litologia classificada e $\bar{L} = [arenito,\ folhelho,\ slurry]$. A classificação é feita para cada litologia $\bar{L}_l$ e a escolhida é aquela onde o valor de probabilidade é maior (maior probabilidade de ocorrer na profundidade). Devido à necessidade dos valores de probabilidade, optamos por utilizar a versão extendida da Equação 4.6:

$$L' = mámixo \left[\frac{P(L = \bar{L}_l) \prod_{k=1}^{m} P(G = \bar{G}_k | L = \bar{L}_l)}{\sum_i P(L = \bar{L}_i) \prod_{k=1}^{m} P(G = \bar{G}_k | L = \bar{L}_i)} \right]_{l=1}. \tag{4.7}$$

A etapa de treinamento T do classificador *naïve* Bayes define as médias μ e os desvios padrão σ de modo que se tenha uma dupla de (μ, σ) para cada litologia e para cada propriedade geofísica (considerando este trabalho, são 9 (μ, σ) para as três litologias e as três propriedades). Estes valores de (μ, σ) são utilizados para estimar, através da FDP,

as probabilidades de $P(G|L)$, de modo que:

$$P(G = \bar{G}_k|L = \bar{L}_l) = \frac{1}{\sigma_{l,k}\sqrt{2\pi}}^{-\frac{1}{2}\frac{(r-\mu_{l,k})^2}{\sigma^2_{l,k}}}. \tag{4.8}$$

Com o intuito de expandir as possibilidades do classificador NB, apresentamos nas seções seguintes algumas estratégias metodológicas para aprimorar o cálculo das probabilidades a priori e as verossimilhanças que compõem a base conceitual do método (i.e., a Equação 4.6).

4.7 Estratégias para Aprimorar o Método *Naïve* Bayes

O classificador *naïve* Bayes, que é a metodologia adotada, é um método supervisionado, e estes usualmente apresentam uma estruturação com duas funções: uma de treinamento (t) e uma de classificação (c). A parte de treinamento é conceitualmente definida como:

$$T = t(\bar{G}_t, L_t), \tag{4.9}$$

em que $\bar{G}_t$ é uma matriz de propriedades com dimensões $o \times m$ onde m é o número de propriedades a serem utilizadas e o é o número de amostras; L_t é um vetor com o elementos constituído por litologias. O sufixo $_t$ é referente a treinamento, e portanto, a matriz $\bar{G}_t$ e a litologia L_t são conhecidas. T é um conjunto de parâmetros a ser definido na etapa de treinamento e que vai ser utilizado durante a classificação conforme:

$$L' = c(\bar{G}_c, T), \tag{4.10}$$

onde $\bar{G}_c$ é uma matriz de propriedades com dimensões $q \times m$ em que q é o número de amostras, e o sufixo $_c$ é referente a classificação, e L' é a litologia classificada.

Nas seções seguintes deste capítulo, apresentamos cinco variações para o método *Naïve* Bayes padrão (STD) que são: *naïve* Bayes sintonizado (TUN), *naïve* Bayes com FDP dada pelo *Kernel density estimation* (KDE), *naïve* Bayes arquitetura (ARC) e *naïve* Bayes com informação estratigráfica prévia (CRC).

4.8 1ª estratégia: Cálculo das verossimilhanças via *Kernel Density Estimation* (KDE)

Alternativamente ao uso da distribuição Gaussiana como função densidade de probabilidade, o KDE é uma estimativa, não paramétrica, da FDP cujo conceito está

associado a Silverman e Jones (1989). Define-se uma estimativa não paramétrica, por aquela que não utiliza parâmetros como, por exemplo, o μ e o σ das distribuições (SPRENT, 1988; CORPORATION, 2014). Portanto, o KDE é uma estimativa que se ajusta primordialmente aos dados, e um de seus propósitos principais seria uma alternativa numérica, contínua do histograma (SCOTT, 2015). Dentre as diversas aplicações do método KDE, este se destaca na análise numérica, no método de diferenças finitas (ROSENBLATT, 1956; TARTER; KRONMAL, 1976), suavização na convolução, aproximação de séries ortogonais (KRONMAL; TARTER, 1968; WATSON, 1969) e na média de histogramas deslocados (*averaging of shifted histograms*) (ROSENBLATT, 1956; SCOTT, 2015). A equação da KDE é dada por (ROSENBLATT, 1956; SCOTT, 2015):

$$f(x) = \frac{1}{nh} \sum_{i=1}^{n} N\left(\frac{x - x_i}{h}\right), \tag{4.11}$$

onde n é o número de amostragens, N é a função kernel, e h é a largura de banda do kernel (ou *bandwidth*). Neste trabalho optamos pelo uso do KDE como uma aproximação numérica para a distribuição dos dados, e utilizamos o algoritmo KDE de Virtanen et al. (2020), cuja formulação é similar a 4.11, e que considera h calculado pela regra de Scott (SCOTT, 2015):

$$h(n, d) = n^{-1/(d+4)}, \tag{4.12}$$

em que d é a dimensão do problema, ou o número de propriedades. A metodologia da variação KDE consiste, portanto, na substituição da FDP gaussiana pelo KDE usando a função gaussiana como kernel.

4.9 2ª estratégia: Utilização do *Naïve* Bayes em Modo *ensemble* (Arquitetura - ARC)

A estratégia arquitetura é baseada no trabalho de Horrocks, Holden e Wedge (2015), onde são considerados dois tipos de arquitetura: singular e conjunta (ou *committee do inglês*). A arquitetura singular é um modo usual onde um único conjunto de dados para treinamento é considerado. Utilizando o método *naïve* Bayes como referência, este considera, unicamente, um conjunto de propriedades geofísicas $\bar{G}$ e um conjunto de litologias $\bar{L}$ que são utilizados nas etapas de treinamento e validação.

A arquitetura conjunta considera que cada poço, que compõe o banco de dados de treinamento, seja uma instância separada de treinamento, ou seja, existe um $\bar{L_w}$ e um

$\bar{G}_w$ onde o sufixo w é referente ao poço. Essencialmente o classificador *naïve* Bayes deve realizar a etapa de treinamento T para cada poço, ou seja:

$$T_w = t(\bar{G}_{t,w}, L_{t,w}), \tag{4.13}$$

e calcular uma classificação L' para cada poço w conforme:

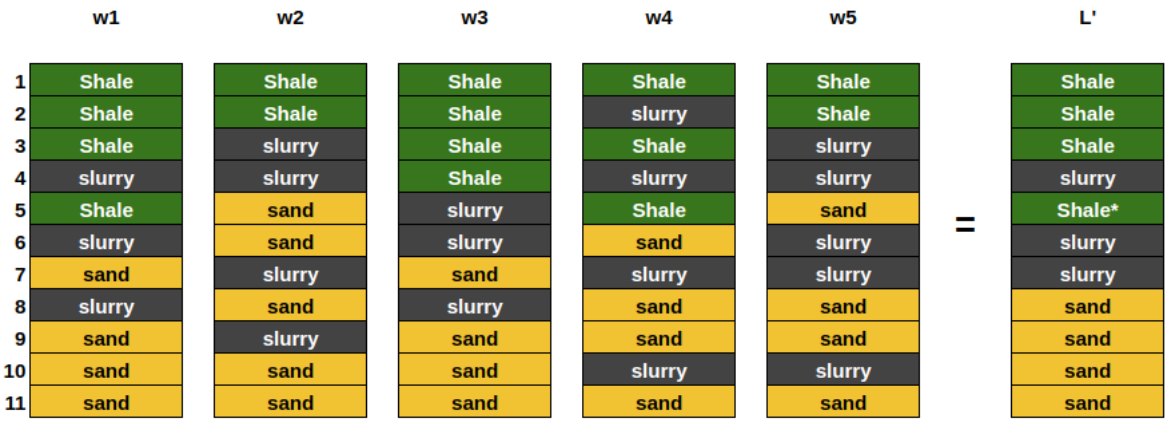
$$L'_w = c(\bar{G}_c, T_w). \tag{4.14}$$

Este processo vai retornar uma quantidade de classificações L'_w que é convertida em uma matriz LW com dimensões $(q \times v)$ onde, novamente, q é o mesmo número em profundidade de $\bar{G}_c$ e v é o número total de poços. Consequentemente, a classificação final L' consiste em determinar a classificação mais comum em profundidade, portanto:

$$L' = mais\ comum\,[LW_c]_{c=1}^{q}\,. \tag{4.15}$$

A Figura 12 apresenta um exemplo esquemático do cálculo do NB ARC. Na profundidade 1, a litologia classificada é L', pois esta é a única existente em todos os poços. Na profundidade 6 a rocha classificada é o *slurry*, visto que esta é a mais comumente encontrada no intervalo. Para o caso de empate como na profundidade 5, o critério é a primeira ocorrência, no caso, o folhelho.

Figura 12 – Exemplo de classificação segundo a estratégia ARC, onde w_{1-5} são os poços e L' a litologia classificada. O $*$ na linha 5 indica um valor de L' obtido pelo critério de desempate.



4.10 3ª estratégia: *Naïve* Bayes Sintonizado (TUN)

Usualmente, os métodos de aprendizado de máquina necessitam de parâmetros que são introduzidos manualmente antes de aplicação do método. Esses parâmetros são

denominados de hiperparâmetros devido a sua independência em relação aos dados, e portanto, considerados em um nível superior se comparados aos parâmetros ajustados durante a etapa de treinamento (AGRAWAL, 2021). Portanto, a etapa de treinamento pode ser redefinida como:

$$T_{tun} = t(\bar{G}_t, L_t, H), \tag{4.16}$$

onde H é relativo aos hiperparâmetros. Portanto, consideramos aqui uma vertente do método *naïve* Bayes onde a *priori* não é definida pelo banco de dados de treinamento, como no método STD, mas um hiperparâmetro a ser ajustado.

4.10.1 Sequência de *Tuning*

Considere que a probabilidade a priori $P(L)$ possa ser definido por um vetor Γ com dimensão n onde seus elementos são as proporções da litologia l; logo, $P(L = l)$ é substituído por Γ_l. Os valores do vetor Γ são gerados de forma aleatória, de modo que o processo de *tuning* direciona esse sorteio aleatório para que o resultado do método apresente o maior valor-f (Equação 3.5).

O algoritmo que realiza o sorteio considera um valor mínimo e um valor máximo para cada proporção de litologia, $\Gamma_{l,min}$ e $\Gamma_{l,max}$ respectivamente, onde o sorteio sempre é realizado de modo que $\Gamma_{l,min} < \Gamma_l \leq \Gamma_{l,max}$, e que $\sum \Gamma = 1$. Consequentemente, o algoritmo inicia considerando $\Gamma_{l,min} = 0$ e $\Gamma_{l,max} = 1$, para qualquer litologia l.

O algoritmo de *tuning* é apresentado no pseudo-código 1, e seu propósito é identificar uma *priori* ótima a partir de *prioris* geradas aleatoriamente, considerando uma convergência

para o maior resultado da soma do valor-f.

Algoritmo 1: NB - TUNING

Entrada: $\bar{G}_t, L_t, \bar{G}_c$,$a$,$d$,$ITE1$,$ITE2$

início
- $\bar{\Gamma} \leftarrow 0$ (inicializa matriz de prioris)
- **para** $j \leftarrow 1, ITE1$ **faça**
  - $\Gamma_l =$ Sorteio $\mid 0 < \Gamma_l < 1$ (para cada litologia l)
  - $\bar{\Gamma} \leftarrow \Gamma_l$ (armazena em $\bar{\Gamma}$ cada ocorrência de Γ_l)
- **fim**
- **para** $j \leftarrow 1, ITE2$ **faça**
  - $C \leftarrow 0$
  - **para** $i \leftarrow 1, ITE1$ **faça**
    - Segregar aleatoriamente $(\bar{G}_t, L_t)$ em $(\bar{G}_{t,a}, L_{t,a})$ e $(\bar{G}_{t,100-a}, L_{t,100-a})$
    - $T = t(\bar{G}_{t,100-a}, L_{t,100-a}, \bar{\Gamma}_i)$.
    - $L'_a = c(\bar{G}_{t,a}, T)$.
    - $C \leftarrow \sum VF(L'_a, L_a)$
  - **fim**
  - $I = idx\ max\ _b(C)$ (índices dos b maiores valores de C)
  - $\bar{\Gamma}_{Mb} = \bar{\Gamma}[I]$ ($\bar{\Gamma}_{Mb}$ contém apenas os valores segundo índice I)
  - $\Gamma_{l,min} = min(\bar{\Gamma}_{l,Mb})$
  - $\Gamma_{l,max} = max(\bar{\Gamma}_{l,Mb})$
  - $\bar{\Gamma} \leftarrow 0$
  - **para** $j \leftarrow 1, ITE1$ **faça**
    - $\Gamma_l =$ Sorteio $\mid \Gamma_{l,min} < \Gamma_l < \Gamma_{l,max}$ (para cada litologia l)
    - $\bar{\Gamma} \leftarrow \Gamma_l$ (armazena em $\bar{\Gamma}$ cada ocorrência de Γ_l)
  - **fim**
- **fim**
- $I = idx\ max(C)$
- $H = melhor(\bar{\Gamma}_{Mb})$ (a priori que apresenta o melhor resultado)
- $T_{tun} = t(\bar{G}_t, L_t, H)$
- $L' = c(\bar{G}_c, T_{tun})$

fim

Saída: L'

.

Para executar o algoritmo 1, é preciso que se tenha uma matriz de parâmetros referente ao treinamento $\bar{G}_t$, um vetor de classes L_t e os parâmetros de classificação $\bar{G}_c$. Os demais valores (a,d,$ITE1$,$ITE2$) não têm relação direta com o dado, e sua função é gerenciar o processo de sintonização. A seguir comentamos mais sobre eles e sobre os processos a que estão associados.

A sintonização assume que não existe um valor conhecido da *priori*, e portanto, esta deve ser inicializada aleatoriamente. Portanto, o algoritmo cria uma listagem aleatória de *prioris* denominada $\bar{\Gamma}$ que vai ser refinada durante todo o processo de *tuning* até que se obtenha a *priori* ótima. O usuário deve, portanto, informar ao algoritmo o número inicial de *prioris* em $ITE1$. Neste trabalho consideramos que $ITE1 = 100$. Na sequência, a variável $ITE2$ é referente a principal iteração da sintonização, ou seja, a porção responsável por otimizar o valor de $\bar{\Gamma}$. Devido ao grande custo computacional deste processo, consideramos que $ITE2 = 10$.

A segunda iteração interna tem o propósito de aplicar o classificador *naïve* Bayes e porções do dado de treinamento, o que pode ser considerado como uma validação cruzada de etapa única. O dado de treinamento é segregado, aleatoriamente, em duas porções, a primeira ($a\%$) utilizada para validação, a segunda ($100 - a\%$) para o treinamento. As litologias classificadas L'_a são comparadas com as litologias L_a através da métrica MVF a ser armazenada na variável C. Optamos por utilizar $a = 40\%$, pois este valor resulta em uma amostragem mais ampla.

Na sequência, ainda interno a primeira iteração, o algoritmo determina os $b\%$ maiores valores de C e, portanto, uma sublistagem de *prioris* ($\bar{\Gamma}_{Mb}$). Dessa sublistagem são selecionados os valores máximos e mínimos de *priori* para cada litologia com o propósito de realizar um novo sorteio de *prioris* $\bar{\Gamma}$. Optamos por $b = 40\%$ pelas mesmas razões discutidas acerca de a.

Após $ITE2$ iterações o algoritmo seleciona, da última sublistagem, a *priori* que obteve o maior valor de VF, aplica o classificador *naïve* Bayes utilizando esta *priori*, e retorna a litologia classificada. A figura Figura 13 apresenta o processo iterativo de sintonização, inciado a partir de várias *prioris* aleatórias.

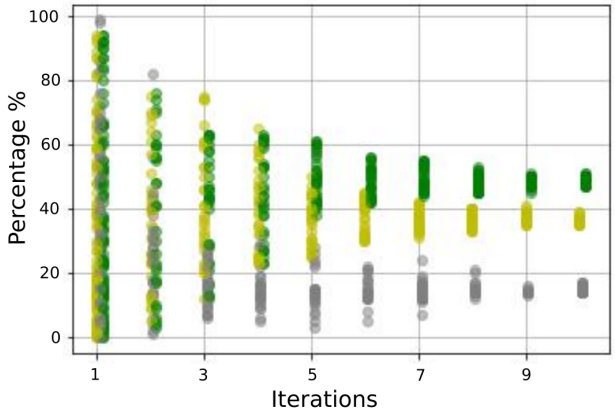


Figura 13 – Convergência das *prioris* considerando M=10 iterações.

4.11 4ª estratégia: *Naïve* Bayes Combinado com Informação Geológica Prévia (CRC)

O CRC é a última estratégia do classificador *naïve* Bayes, e possui semelhança com o ARC quanto a sua estruturação. O CRC considera zonas em profundidade referentes aos estágios turbidíticos crc-1, crc-2 e crc-3 (FREIRE et al., 2020; SILVA, 2020; SILVA et al., 2018). Pela perspectiva classificatória, estas zonas indicam uma homogeneidade local referentes à sua gênese e, consequentemente, podem ser classificadas de modo independente tal como apresentado na seção 4.9. Portanto, considere um conjunto CR de modo que seu espaço amostral é definido por $\{1, 2, 3\}$, e que cr é uma zona genérica de CR, logo, o treinamento pode ser definido como:

$$T_{cr} = t(\bar{G}_{t,cr}, L_{t,cr}, V), \tag{4.17}$$

onde $\bar{G}_{t,cr}$ é a matriz de propriedades e $L_{t,cr}$ é a litologia, considerando apenas a zona cr em profundidade, e V é a verossimilhança (calculada através da FDP gaussiana). A classificação é definida por:

$$L'_{cr} = c(\bar{G_{c,cr}}, T_{cr}), \tag{4.18}$$

em que $\bar{G_{c,cr}}$ é a matriz de propriedades do poço a ser classificado considerando a zona cr. Nesta estratégia, a verossimilhança é calculada para o campo como um todo . A litologia final L' é obtida:

$$L' = [L_{CR_i}]_{i=1}^{3}. \tag{4.19}$$

Portanto, para utilizar a alternativa CRC é necessário que o poço a ser classificado já esteja zonado, o que pode ser realizado através da interpretação do perfil GR, comumente presente nos poços, mesmo nos mais antigos. O método CRC aplicado neste trabalho considera um zoneamento associado ao fluxo turbidítico, e portanto, está associado à interpretação geológica restrita para esta bacia, no entanto, este abre um precedente para o uso de outros tipos de zoneamento com correlação estratigráfica nos poços.

5 Resultados

A análise dos resultados compreende duas seções principais: uma para a validação e outra para a aplicação real. A seção de validação se inicia com uma análise geral das classificações obtidas por cada estratégica para cada poço que compõe o banco de dados de treinamento, através das dispersões/distribuições e das tabelas de valor-f e erro. Posteriormente, segue-se uma análise individual para cada estratégia, que é concluída com uma análise comparativa entre todos os resultados. A seção de aplicação real consiste na classificação de eletrofácies em um poço sem descrição litológica utilizando o *naïve* Bayes e as estratégias apresentadas no Capítulo 4

5.1 Distribuição e Análise dos Dados de Treinamento para o Campo de Massapê

O cálculo da verossimilhançaa $P(G|L)$ foi efetuado à partir dos dados de 12 poços do Campo de Massapê, Bacia do Recôncavo. A estruturação destes dados é apresentada na Figura 14 onde expomos os gráficos de dispersão e os histogramas do banco de dados por litologia. Neste caso, os perfis GR, DT e log(ILD) são apresentados aos pares, cujo objetivo é verificar o comportamento da distribuição dos dados interpretados a partir do perfil DRDN (FREIRE et al., 2020), e como estes se correlacionam. Os resultados tendem a ser melhores quando existe alguma separabilidade espacial entre os dados referentes a cada litofácies. Teoricamente, o método NB (e correlatos) desconsideram a dispersão, pois tem como premissa o principio da independência entre propriedades (seção 4.5), mas a distribuição evidencia uma correlação incipiente entre as propriedades.

A princípio observamos uma boa separação entre arenitos e folhelhos principalmente quando o perfil GR é considerado, conforme pode ser observado na Figura 14 a). No entanto, há uma considerável sobreposição entre os arenitos e os slurries, o que implica em maior dificuldade na classificação dessas eletrofácies. Outro ponto mencionável é que os gráficos de dispersão que não envolvem o perfil GR apresentam uma maior sobreposição nas litologias interpretadas, tornando-o essencial na classificação de arenitos e folhelhos no Campo de Massapê. A realização desta análise prévia é de suma importância visto que esta indicam que haverá uma boa classificação para o folhelho, e razoável para o arenito e *slurry*. Os gráficos da Figura 14, apresentam as dispersões entre as propriedades (estas condicionadas às litologias), assim como, as distribuições para cada propriedade/litologia. Nas distribuições, as curvas em vermelho são referentes a FDP Gaussiana, utilizada nas estratégias STD, ARC, TUN e CRC; enquanto que as azuis foram as FDP's utilizadas

pela estratégia KDE, onde o ajuste às distribuições é mais preciso.

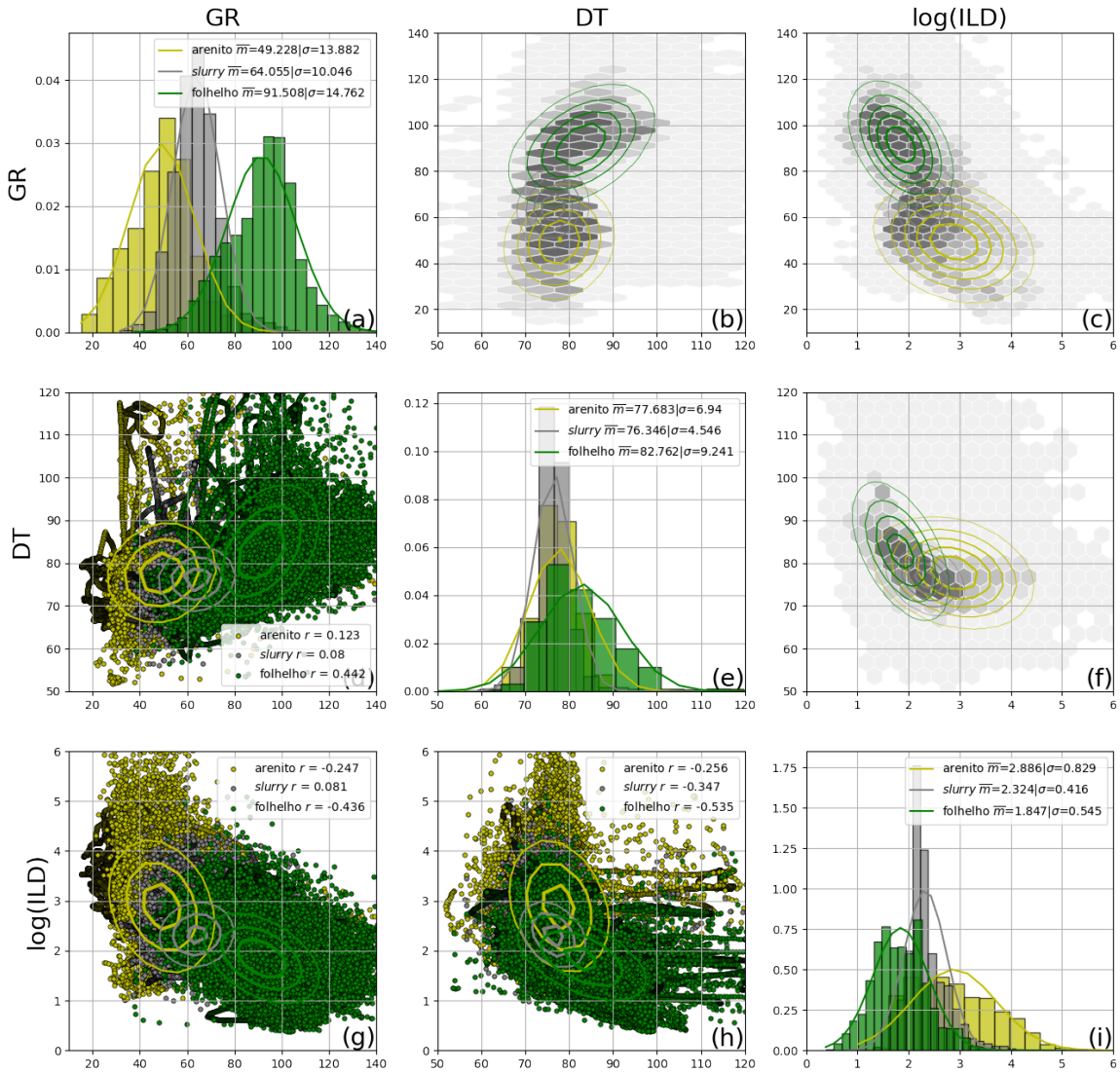


Figura 14 – Gráficos de dispersão e distribuição do banco de dados de treinamento; Letras a, e, i apresentam as verossimilhanças por litologia, tanto ajustada via distribuição normal (curva nas cores da litologia), assim como valores da média $\bar{m}$ e do desvio padrão σ para cada litologia; a triangulação esquerda, referente as letras d, g, h, são os gráficos de dispersão para cada litologia, apresentando também o coeficiente de correlação r e gaussianas bidimensionais para cada litologia (curvas de nível nas cores da litologia); A triangulação direita, referente as letras b, c, f, apresenta o gráfico de compartimentação hexagonal, e as gaussianas bidimensionais para cada litologia com sua respectiva coloração.

5.1.1 Análise do Campo de Massapê: Erro e valor-f

De um modo geral, todas as estratégias apresentam resultados bem similares. Os erros médios oscilaram na faixa de 20% e a estratégia que apresentou o menor erro foi a CRC, com erro médio de 20.315%, sendo o maior erro médio foi referente ao STD com 22.038%. A Figura 15 apresenta uma tabela com os erros para cada poço utilizado

no Campo de Massapê, onde podemos observar que o erro tende a ser similar entre as estratégias, e distinto entre os poços.

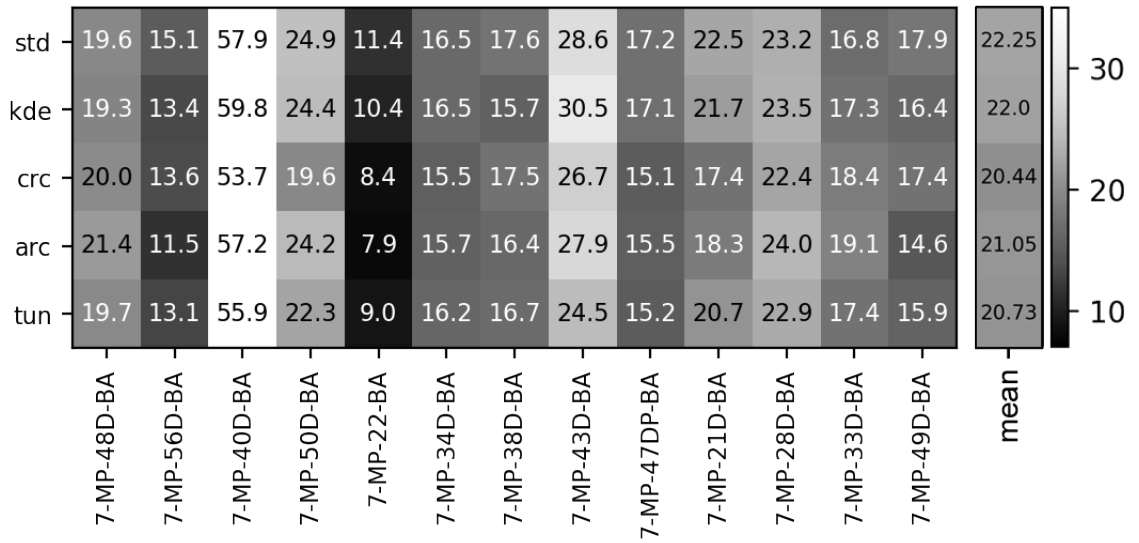


Figura 15 – Tabela com os valores dos erros para cada poço, acrescido da média para cada estratégia.

Considerando a média do valor-F, a melhor estratégia foi a CRC com resultado de 0,69 e a pior foi a TUN com resultado de 0,673. A Figura 16 expõe os valores do valor-f referentes a cada poço onde observamos, novamente, que os valores tendem a ser parecidos entre as estratégias e com variação entre os poços.

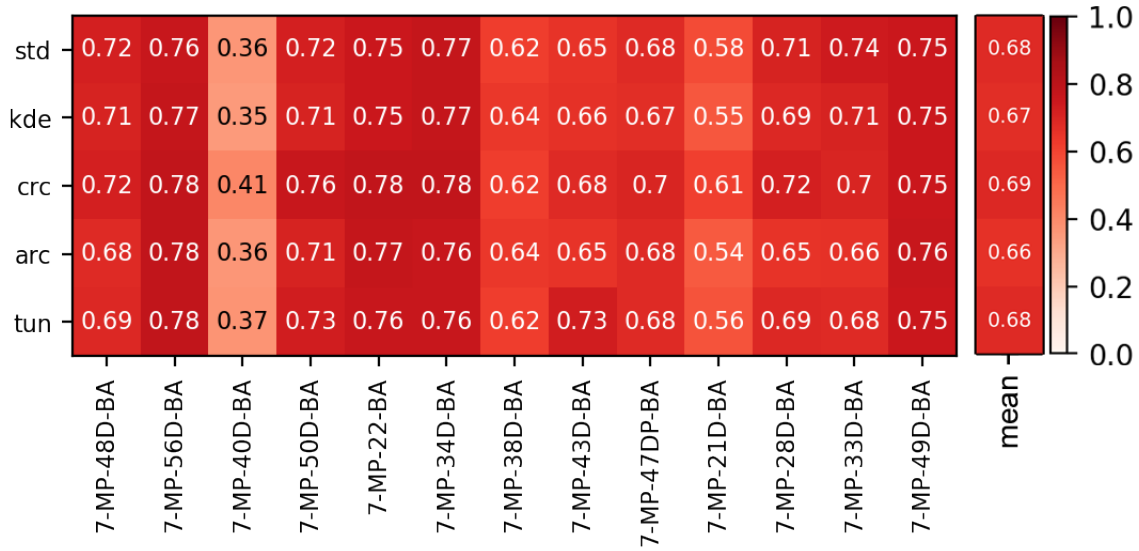


Figura 16 – Tabela com os valores de MVF para cada poço, acrescido da média para cada estratégia.

Considerando os resultados apresentados, a estratégia CRC é aquela que se apresenta com maior eficácia na classificação de dados de poços, seguida diretamente pela ARC, com resultados bem promissores. Apesar de não demonstrar uma melhora substancial em

relação à estratégia tradicional STD, e de sempre apresentar valores similares, a CRC apresentou o melhor resultado também na contagem poço a poço, ou seja, a CRC tem o maior valor-f em oito dos 13 poços, e o menor erro em quatro dos 13 poços. Deste modo, interpretamos que existe uma distinção regional considerável entre os poços, devido a variação do erro nas tabelas, e pelos resultados consideráveis do método ARC, mas que também existe uma homogeneidade estratigráfica proeminente nas zonas referentes a estratégia CRC. Dessa forma, definimos o poço 7MP 50 BA como o de validação das estratégias apresentadas neste trabalho, especialmente por apresentar as maiores diferenças entre os valores-F computados. Isso reforça o fato do zoneamento ser um fator diferencial para refinar o uso do classificador NB.

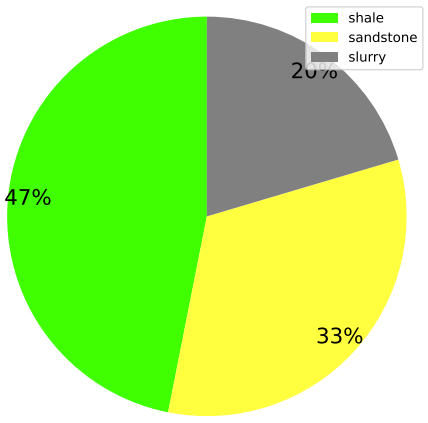
5.2 Poço de Validação :7-MP-50D-BA

Nesta seção, uma análise mais específica é apresentada considerando um único poço. O propósito é avaliar as estratégias aplicadas neste trabalho de modo mais usual à análise petrofísica. Portanto, foi escolhido o poço 7-MP-50D-BA, pois este apresenta uma diferença considerável da CRC para com as demais estratégias (tecnicamente o 7-MP-40D-BA seria aquele com a maior diferença, mas devido ao grande erro associado a este poço, não o consideramos como representativo do campo).

5.2.1 Avaliação da Classificação via STD

Para esta estratégia, consideramos como *priori* a proporção de litologias tomadas para o campo inteiro (com exceção ao poço de validação), e está representada na figura Figura 17. As verossimilhanças desta estratégia foram calculadas com base nas FDP's gaussianas.

Figura 17 – Gráficos em pizza para todos os 12 poços do Campo de Massapê.



O erro aproximado desta estratégia foi de 25% (Figura 18 - f), e está mais evidente na profundidade de 2650 m. Acima deste ponto, na porção correspondente aos espessos

pacotes de folhelho, a ocorrência de erro é menor, principalmente se comparado com a porção mais profunda do poço. Este comportamento estaria associado a facilidade do método em diferenciar a porção lacustre da turbidítica (folhelho/não folhelho) e a uma maior dificuldade em diferenciar as fácies turbidíticas (arenito/*slurry*), estas comumente presentes na porção inferior do intervalo. Esta tendência esta presente tanto no perfil classificado, quanto no perfil de erros, evidenciando, segundo este último, que os erros são mais comuns na porção arenosa, e que geralmente se apresentam em blocos. Observa-se também que quando o perfil GR está alto, o modelo tende a classificar as rochas como folhelho, e quando está baixo, como arenito. Um comportamento similar (porém inverso), pode ser observado no perfil log(ILD). Quando ambos os perfis GR e log(ILD) são baixos, ou seja, nas porções do arenito com baixa saturação de hidrocarbonetos, os erros tendem a ser maiores. Este é justamente o caso observado nos arredores da profundidade 2650 m, onde o erro é bem pronunciado em dois blocos no que seriam arenitos erroneamente interpretado como *slurry*. O perfil de probabilidades evidencia que o método é bem assertivo em definir o folhelho em relação às demais litologias, enquanto que tende a ser irresoluto na definição de arenitos e *slurries*. Uma tendência observada é que, para a classificação de arenitos, o método costuma ser mais assertivo, se comparado às rochas classificadas como *slurries*, sempre existe uma probabilidade razoável para que a rocha no intervalo seja um arenito Figura 18.

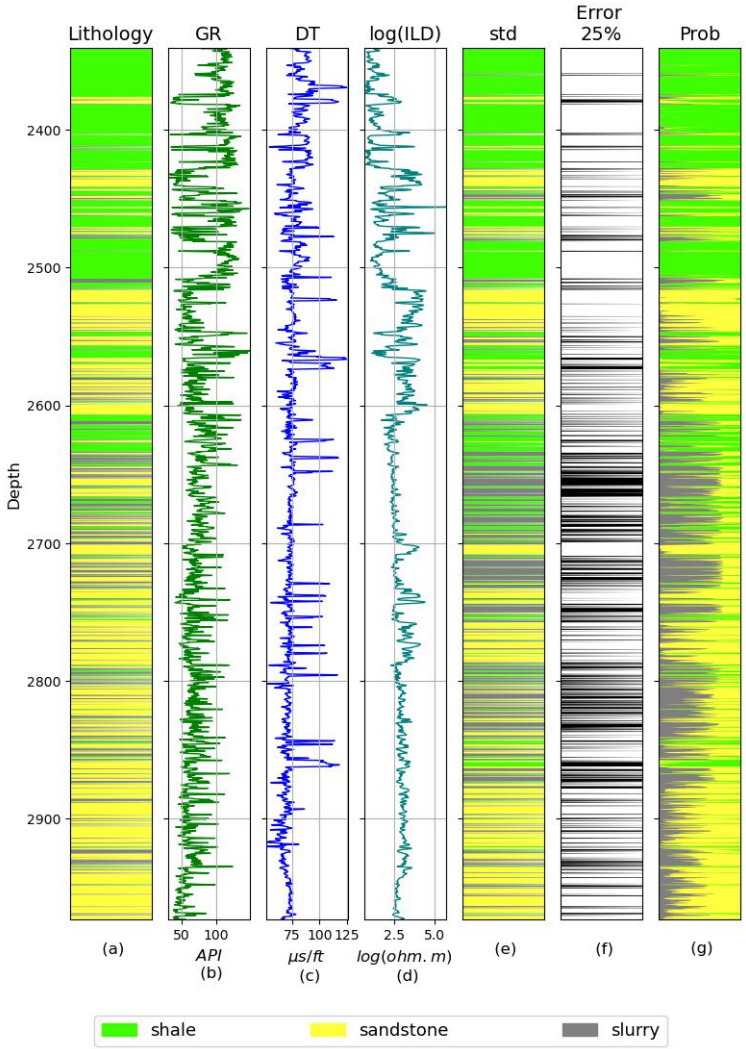
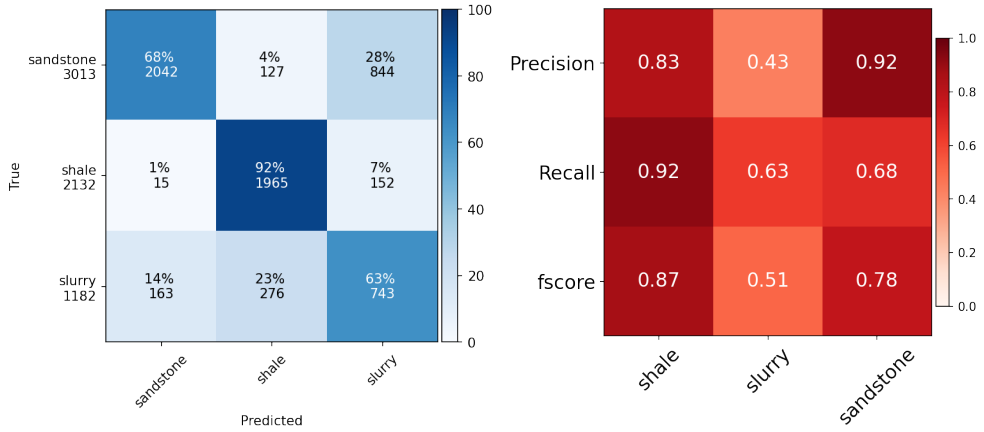


Figura 18 – Classificação do poço de validação via estratégia STD onde: (a) litologia interpretada; (b) perfil GR; (c) perfil DT; (d) perfil log(ILD); (e) litologia classificada; (f) perfil de erros; (g) perfil de Probabilidades.

A matriz de confusão relativa ao STD (Figura 19 (a)) salienta a efetividade do classificador na identificação de folhelhos, mas que apresenta alguma dificuldade em classificar arenitos e *slurries*, de modo que uma porção considerável dos arenitos (28%) foi confundida com o *slurry*. A precisão e o valor-f indicam a dificuldade do método em identificar o *slurry*, enquanto que os valores altos da revocação no *slurry* estariam associados a baixa quantidade de falsos positivos, referente a uma tendência do método em confundir arenito com slurry (Figura 19 (b)).



(a) (b)

Figura 19 – Matriz de confusão (a) e tabela com valores de precisão, revocação e valor-f (b) para o *naïve* Bayes STD

5.2.2 Avaliação da Classificação via KDE

Esta estratégia considera, como *priori*, a proporção de litologia no campo inteiro, ou seja, a mesma da STD. A grande mudança está relacionada ao uso da distribuição homônima (*kernel density estimation*) apresentada na Figura 14 letras (a), (e) e (i) (curva azul).

A estratégia KDE apresentou resultados bem similares em comparação com o STD, conforme pode ser observado pelos valores bem próximos do erro (Figura 20 (f)). Uma maior distinção está bem pronunciada próxima a profundidade de 2800 m, onde observamos duas camadas finas de *slurry* em meio a uma zona de intercalações entre as três litologias. Observamos também uma maior tendência em valorizar o folhelho, algo que pode ser observado pela coloração mais esverdeada do perfil da Figura 20 (e).

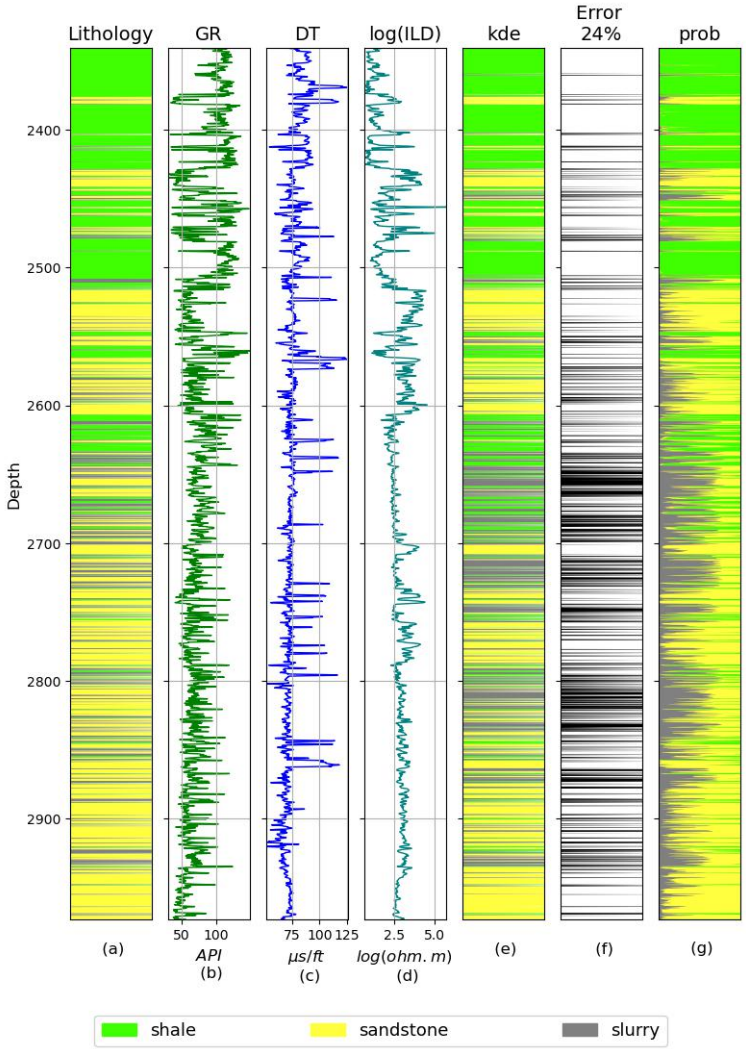
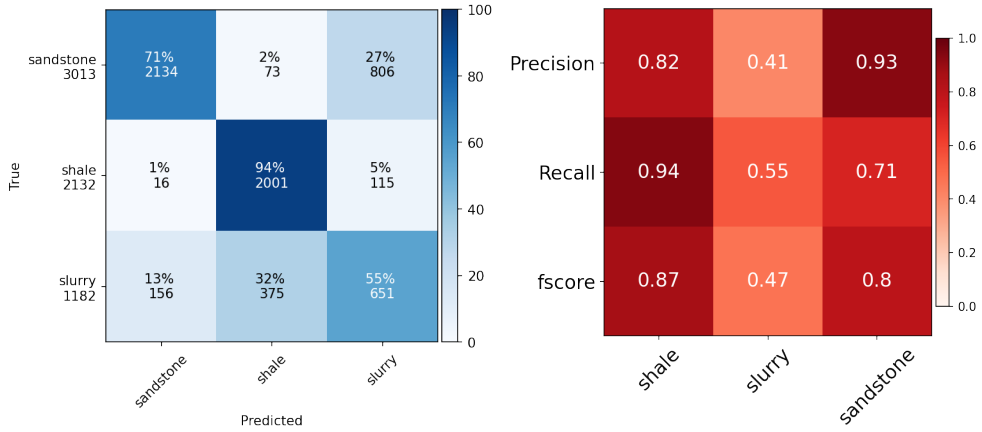


Figura 20 – Classificação do poço de validação via estratégia KDE, onde: (a) litologia interpretada; (b) perfil GR; (c) perfil DT; (d) perfil log(ILD); (e) litologia classificada; (f) perfil de erros; (g) perfil de Probabilidades.

Esta tendência (de privilegiar o folhelho) é bem evidente na matriz de confusão (Figura 21 (a)), onde podemos observar que uma porção considerável do *slurry* é erroneamente classificada como folhelho. Salientamos também que esta estratégia resultou em um aprimoramento na classificação do arenito. Adicionalmente, a KDE apresentou uma revocação relativamente maior para o arenito (Figura 21 (b)), indicando uma maior quantidade de verdadeiros positivos (TP). De modo geral o KDE privilegiou as litologias que são quantitativamente dominantes (arenito e folhelho) em detrimento do *slurry*.



(a) (b)

Figura 21 – Matriz de confusão (a) e tabela com valores de precisão, revocação e valor-f (b) para a estratégia KDE

5.2.3 Avaliação da Classificação via TUN

A estratégia TUN considera a distribuição gaussiana para o campo inteiro (excluindo o poço de validação), e uma *priori* aleatória otimizada através de 10 iterações. O resultado destas iterações é apresentado no gráfico em pizza na Figura 22.

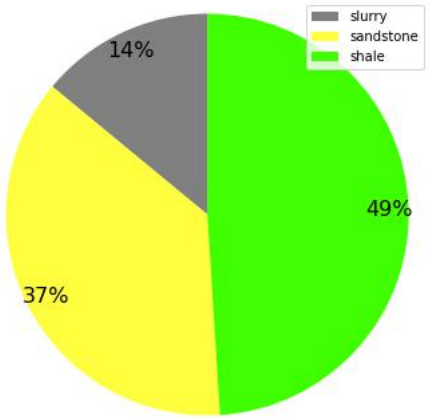


Figura 22 – *Priori* ótima calculada após 10 iterações.

Esta estratégia demonstra uma tendência em valorizar mais o arenito em relação ao *slurry*, algo que é bem perceptível ao se observar o perfil da Figura 23 (e), próximo a profundidade de 2800 m. Esta estratégia modifica apenas a *priori* do método, algo que está bem evidente no perfil de probabilidades (Figura 23 (g)) que, comparado ao STD, apresenta uma probabilidade menor para o *slurry*.

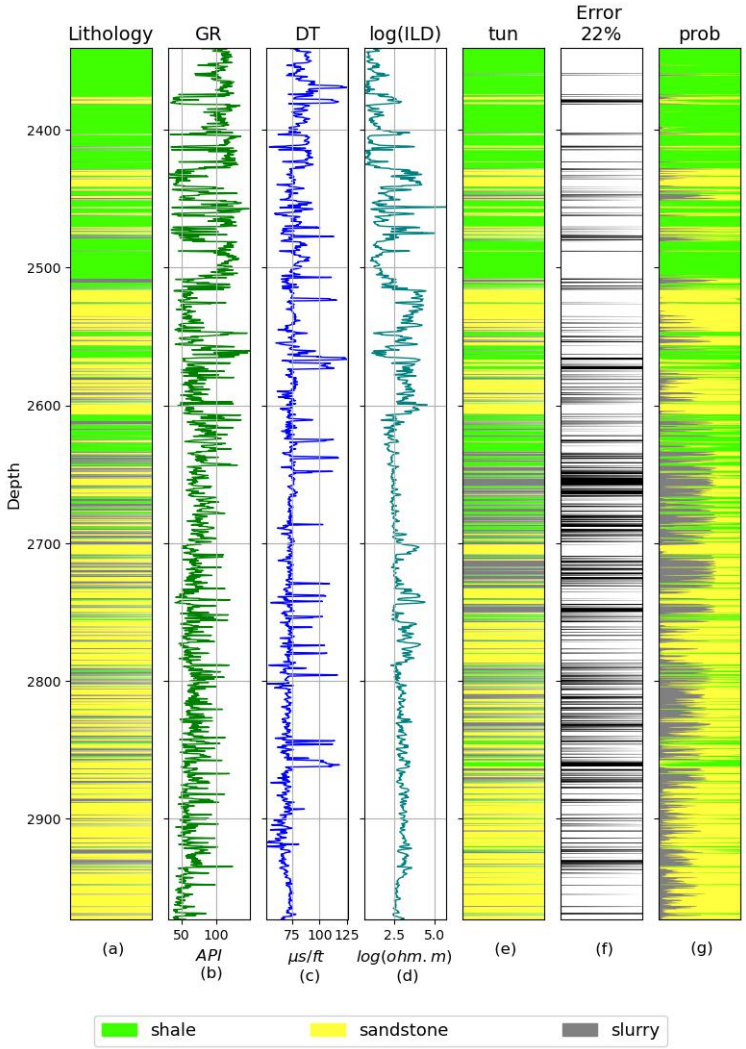
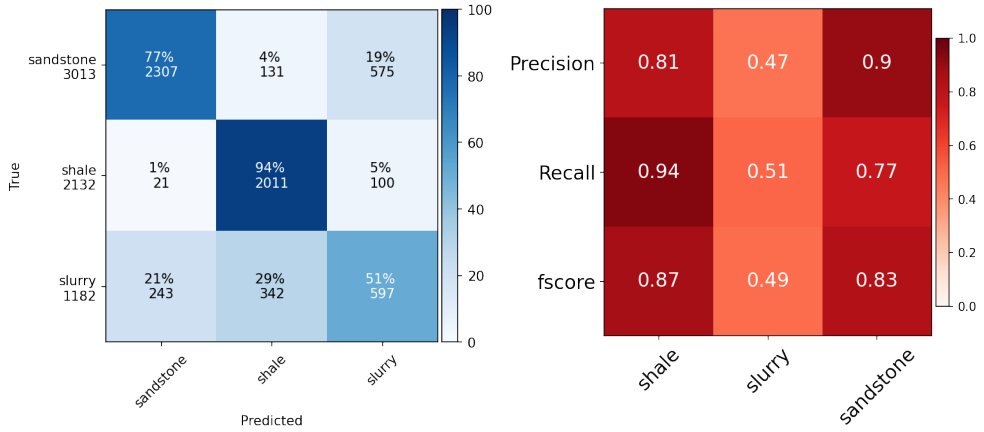


Figura 23 – Classificação do poço de validação via estratégia TUN, onde: (a) litologia interpretada; (b) perfil GR; (c) perfil DT; (d) perfil log(ILD); (e) litologia classificada; (f) perfil de erros; (g) perfil de Probabilidades.

A análise da matriz de confusão demonstra que esta estratégia tende a melhorar a classificação para o arenito, enquanto que piora a dos slurries, ou seja, a estratégia prioriza o arenito na classificação (Figura 24 (a)). Este processo é evidenciado pela métrica de revocação com incremento no arenito, e a redução para o *slurry* (Figura 24 (b)) e pode ser justificado como uma maior quantidade de *slurry* sendo classificada como arenito (falsos positivos do arenito).



(a) (b)

Figura 24 – Matriz de confusão (a) e tabela com valores de precisão, revocação e valor-f (b) para a estratégia TUN

5.2.4 Avaliação da Classificação via ARC

A estratégia ARC considera uma *priori* e uma verossimilhança para cada poço do campo de massapê, desconsiderando o poço 7-MP-50D-BA. Devido a complexidade em apresentar estes dados, optou-se por omití-los nesta seção.

Esta estratégia apresenta uma tendência em supervalorizar o folhelho, o que pode ser observado nas duas camadas prominentes próximas à profundidade de 2850 m erroneamente classificadas como folhelho (Figura 23 (e)). O perfil de probabilidades indica que, neste intervalo, a litologia poderia ser classificada como arenito ou folhelho, o que não é usualmente observado (Figura 23 (e)).

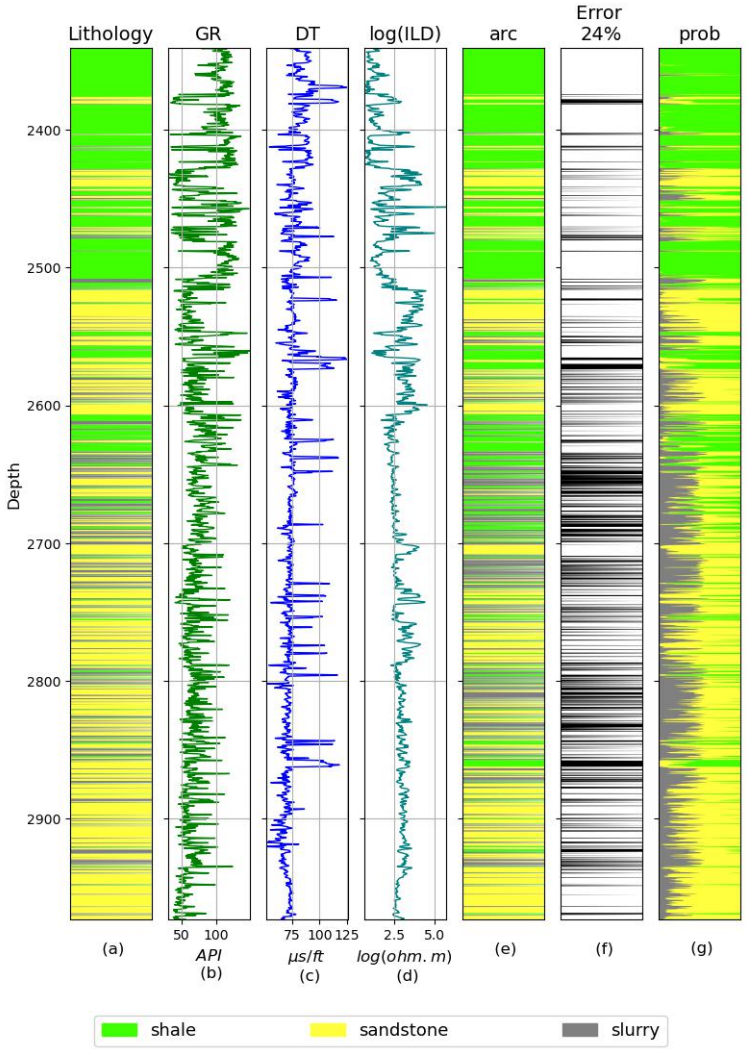
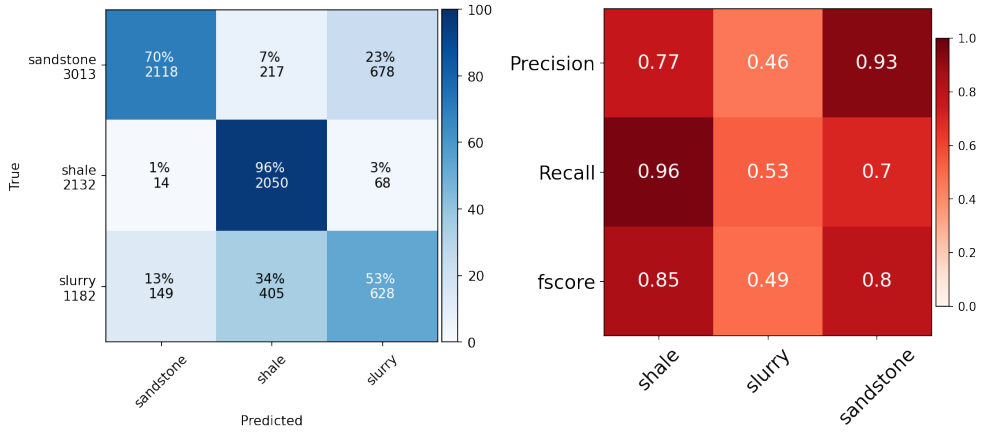


Figura 25 – Classificação do poço de validação via estratégia ARC, onde: (a) litologia interpretada; (b) perfil GR; (c) perfil DT; (d) perfil log(ILD); (e) litologia classificada; (f) perfil de erros; (g) perfil de Probabilidades.

A análise da matriz de confusão demonstra que esta estratégia tende a classificar uma quantidade considerável de *slurry* como folhelho (falsos positivos do folhelho), enquanto que melhora (sutilmente) a classificação dos arenitos (Figura 24 (a)). Este processo é evidenciado também pela redução da precisão na classificação do folhelho, assim como, pelo aumento na revocação do arenito (Figura 24 (b)).

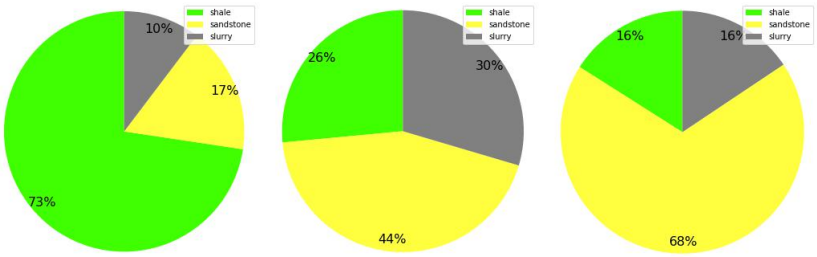


(a) (b)

Figura 26 – Matriz de confusão (a) e tabela com valores de precisão, revocação e valor-f (b) para a estratégia ARC

5.2.5 Avaliação da classificação via CRC

A estratégia CRC considera um zoneamento em profundidade relativos a zonas produtoras. Os intervalos de profundidade destas zonas, para o poço 7-MP-50D-BA, são: CRC 1 [2341 a 2496]; CRC 2 [2496 a 2657]; e CRC 3 [2657 a 2973]. A Figura 27 apresenta as *prioris* utilizadas para o cálculo em cada zona.



(a) CRC 1 (b) CRC 2 (c) CRC 3

Figura 27 – Gráficos em pizza mostrando as prioris para cada zona considerando o poço de validação 7-MP-50D-BA.

Em termos gerais, a estratégia CRC apresentou resultados bem similares se comparados a aplicação STD, principalmente quanto ao erro concentrado próximo à 2650 m e uma maior frequência de erros abaixo deste ponto. O perfil da Figura 28 (h) indica que a presença de três zonas referentes ao fluxo turbidítico, com transições nas profundidades aproximadas de 2500 e 2650 m, esta última coincidindo com a região de maior erro nas classificações. Esta estratégia apresenta um erro de 20% e uma forte tendência em supervalorizar os arenitos na região relativa à profundidade de 2657 a 2973 (CRC 3), algo que pode ser observado também no perfil de probabilidade, com uma menor probabilidade

(em geral) de se classificar slurries.

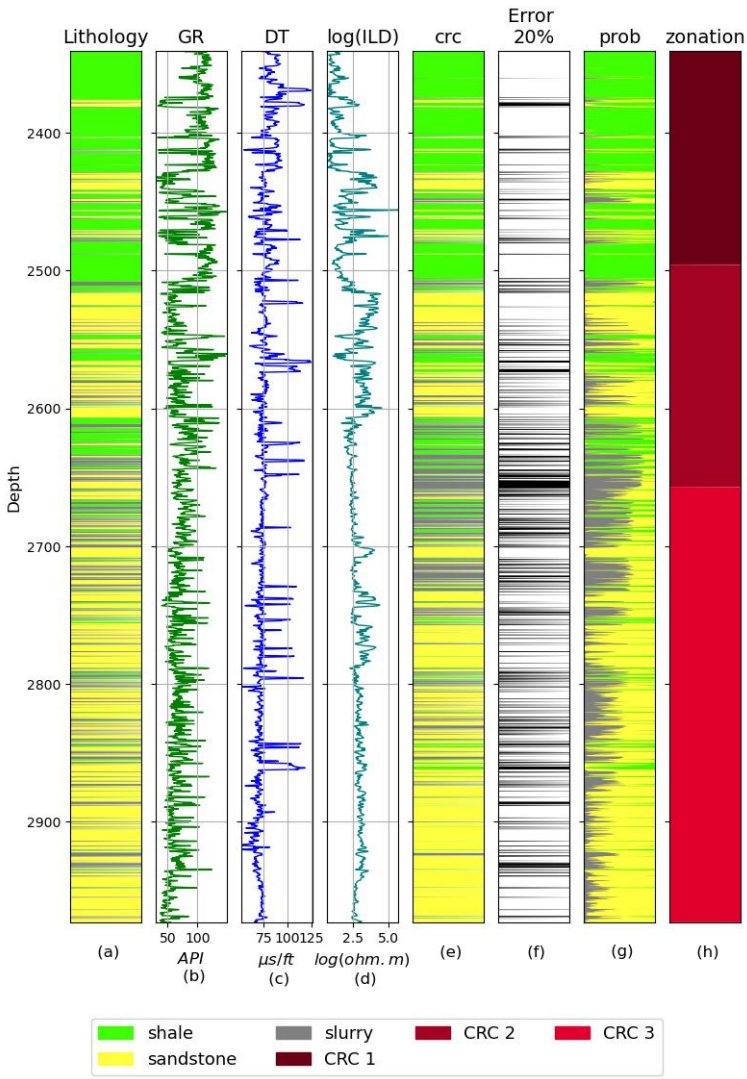
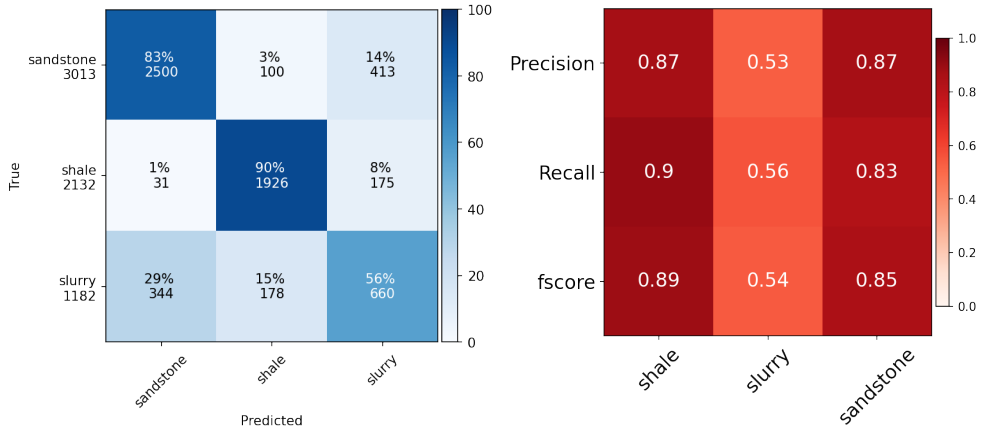


Figura 28 – Classificação do poço de validação via estratégia CRC, onde: (a) litologia interpretada; (b) perfil GR; (c) perfil DT; (d) perfil log(ILD); (e) litologia classificada; (f) perfil de erros; (g) perfil de Probabilidades.

A estratégia CRC, conforme pode ser observado pela matriz de confusão da Figura 28 (a), apresenta uma forte tendência em criar falsos positivos para o arenito, devido à *priori* que é mais favorável a esta litologia no intervalo. Este processo resulta em uma redução na precisão da classificação dos arenitos, mas aumenta a revocação (Figura 28 (b)). Considerando o valor-f, todas as litologias apresentaram um incremento em relação ao STD, com destaque para o arenito.



(a) (b)

Figura 29 – Matriz de confusão (a) e tabela com valores de precisão, revocação e valor-f (b) para a estratégia KDE

Por fim, a Figura 30 apresentada compreende as classificações efetuadas para cada estratégia. Inicialmente, observa-se que os métodos são bem similares, em particular para a porção superior do poço. Salientamos também um grande destaque para a CRC, principalmente abaixo da profundidade de 2650 m, onde esta estratégia apresentou uma forte tendência em supervalorizar o arenito, processo este que faz com que esta se aproxime mais da litologia original (Figura 30 (a)). Nenhum método conseguiu identificar corretamente as duas camadas de arenito próximas a 2650 m muito provavelmente devido ao baixo valor do log(ILD) neste intervalo.

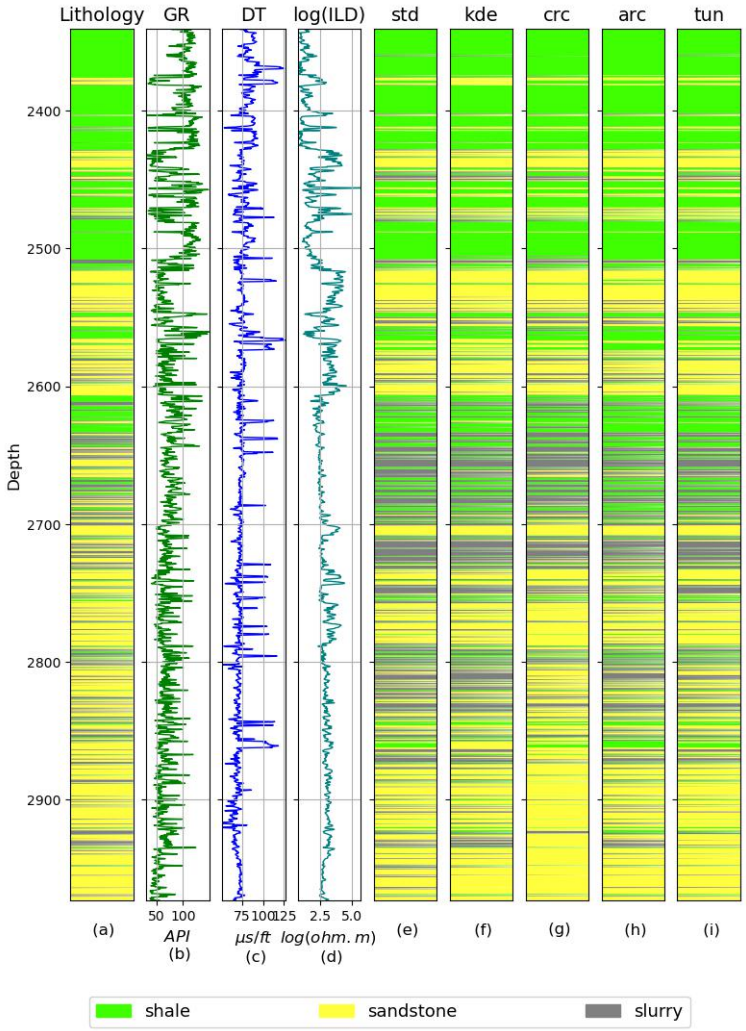


Figura 30 – Classificações do poço de validação usando todas as estratégias, onde: (a) litologia interpretada; (b) perfil GR; (c) perfil DT; (d) perfil log(ILD); (e) estratégia STD; (f) estratégia KDE; (g) estratégia CRC; (h) estratégia ARC; (i) estratégia TUN;

5.2.6 Poço de classificação

A Figura 31 apresenta uma junção de todos os resultados do método *naïve* Bayes e estratégias correlatas para o poço 5-BRSA-365-BA. A análise comparativa mostra uma discrepância para a classificação do arenito, onde as estratégias STD, CRC e TUN *naïve* apresentam um arenito bem definido, blocado, com formato similar ao encontrado nos outros poços; enquanto que a KDE e a ARC apresentaram camadas bem finas de arenitos, em baixas quantidades, apresentando algum agrupamento nas regiões onde as propriedades apresentam formas distintas (por volta de 1090, 1170 e 1250 m). Considerando os resultados podemos interpretar que o folhelho é a eletrofácie predominante neste poço, que existe algum *slurry* disperso e bem definido em todos os métodos, e que, muito provavelmente existem duas camadas de arenito próximo das profundidades de 1100 e 1250 metros. Em essência, o não reconhecimento destas fácies pelas estratégias KDE e ARC estaria associado

a baixa do perfil log(ILD), devido a ausência de hidrocarbonetos.

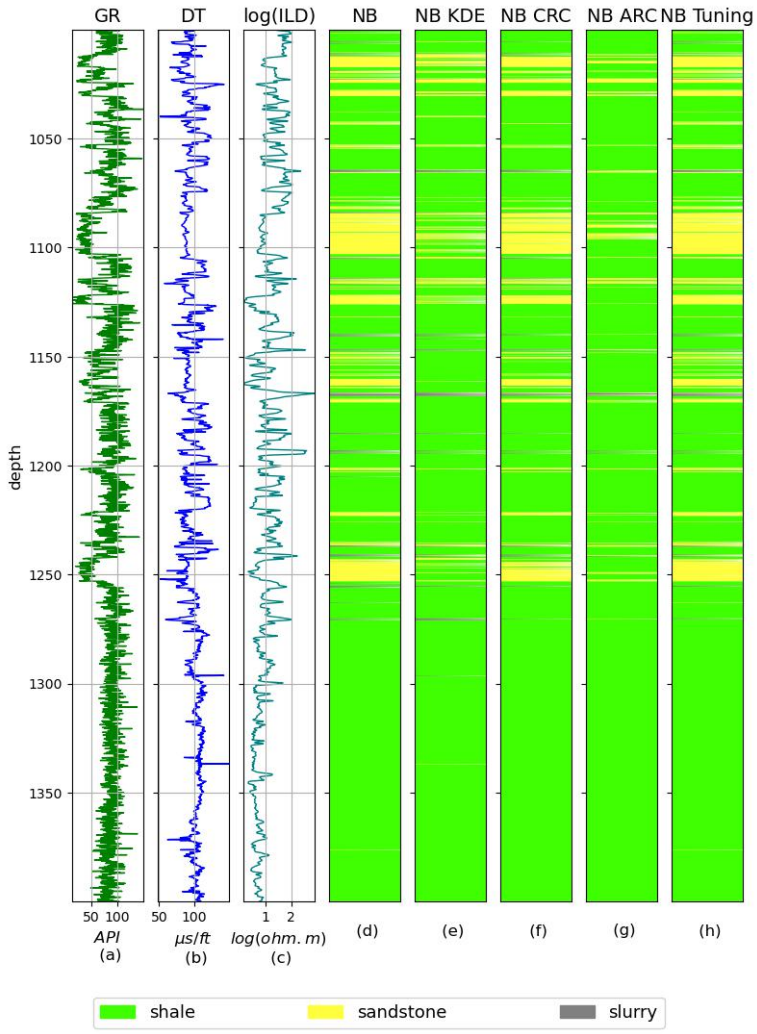


Figura 31 – Aplicação das estratégias ao poço de classificação 5-BRSA-365-BA, onde: (a) perfil GR; (b) perfil DT; (c) perfil log(ILD); (d) Litologia classificada através da estratégia *naïve* Bayes; (e) Litologia classificada através da estratégia KDE; (f) Litologia classificada através da estratégia *naïve* Bayes faciológico; (g) Litologia classificada através da estratégia *naïve* Bayes arquitetura; (h) Litologia classificada através da estratégia *naïve* Bayes tuning

A análise individual mostra que no perfil GR as litologias tem um padrão esperado segundo o regime de fluxo que domina a deposição, onde as eletrofácies referentes a uma granulometria mais fina apresentam um valor maior, enquanto que as mais grosseiras tem um valor menor. No perfil DT (sônico), observa-se uma forte sobreposição entre o arenito e o *slurry*, onde este último apresenta os menores valores de média, enquanto que a distribuição do folhelho é mais distintas. Por fim, para o perfil log(ILD) existe uma sobreposição considerável entre as eletrofácies, mas as distribuições ainda são bem definidas.

6 Conclusões

Neste trabalho foram utilizadas cinco estratégias relacionadas ao classificador *naïve* Bayes com o intuito de promover uma utilização mais efetiva da técnica de AM. Estas estratégias foram utilizadas no problema da classificação de eletrofácies, a partir de perfis geofísicos de 12 poços referentes ao Campo de Massapê, Bacia do Recôncavo. Foram feitas duas avaliações: para o campo como um todo, e para um perfil representativo referente à melhor das estratégias (CRC). As métricas para avaliação foram: perfil de erros, precisão, revocação e valor-f. Posteriormente, estas estratégias foram utilizadas na classificação do poço 5-BRSA-365-BA, este desprovido de perfil litológico.

Os resultados foram promissores quanto ao uso destas estratégias na predição de litologias utilizando os perfis GR, DT e log(ILD), considerando que apenas um dos perfis (GR) é usualmente utilizado na classificação de rochas. É importante salientar que o perfil log(ILD) é sensível ao óleo presente dentro dos poros da rocha, e portanto, a caracterização do arenito e do *slurry* está condicionada à saturação de óleo. Como a etapa de treinamento foi realizada considerando rochas que contém hidrocarbonetos, todas as estratégias tendem a apresentar erros pronunciados em regiões com baixa saturação de óleo. Outro ponto que deve ser considerado é que o perfil log(ILD) apresenta uma tendência resistiva em torno do reservatório, e que essa tendência não foi removida durante a etapa de pré-processamento. A validação indicou que as metodologias tendem a apresentar resultados similares entre si, e que a maior distinção ocorre geograficamente, isto é, cada poço apresenta um comportamento particular. Dentre todas as estratégias, aquela que mais se destacou foi a CRC que é referente a zonas em profundidade associadas ao processo turbidítico.

Uma classificação automática foi realizada para o poço 5-BRSA-365-BA, este sem perfil litológico interpretado. Este poço era o único onde os perfis GR, DT e log(ILD) estavam presentes e que só possuía o RHOB para a classificação. Todas as estratégias foram aplicadas e os resultados foram bem distintos, confrontando a perspectiva vigente de que as estratégias (em particular a KDE e a STD) apresentam resultados similares. O poço 5-BRSA-365-BA, em conjunto com o perfil log(ILD), aparenta enaltecer uma tendência particular às estratégias KDE e ARC, que é a sensibilidade aos dados. Como o treinamento foi aplicado em uma região com uma presença considerável de hidrocarbonetos, a classificação torna-se tendenciosa para hidrocarbonetos. Essa tendência estaria sendo mitigada nas estratégias STD, TUN e CRC devido à utilização da gaussiana como FDP (reforçamos que as estratégias TUN e CRC atuam majoritariamente na *priori*).

Apesar do resultado duvidoso da estratégia ARC na classificação do poço 5-BRSA-

365-BA, seus resultados foram significativos na classificação (segundo melhor método), e portanto, pode ser utilizada em conjunto com outras estratégias. Salienta-se que o volume de amostras também pode ter sido um fator relevante na determinação da *priori* do TUN, e portanto, sugere-se uma análise com uma amostragem constante para cada litologia. A TUN foi aplicada considerando apenas a *priori*, mas esta estratégia pode ser aplicada também na *bandwidth* da verossimilhança (ou nos dois).

7 APÊNDICE

Esta seção é dedicada ao artigo intitulado "*Analysis of alternative strategies applied to Naive-Bayes classifier into the recognition of electrofacies: Application in well-log data at Recôncavo Basin, North-East Brazil*"submetido à revista *Journal of Petroleum Science and Engineering* no dia 31/05/2022. O texto do artigo conta com a autoria de Mário Martins Ramos, Rodrigo Bijani, Fernando Vizeu, Wagner Lupinacci e Fernando Freire. O abstract do artigo doi removido deste apêndice por motivos de revisão do trabalho junto ao corpo editorial da revista.

7.1 Introduction

Machine learning (ML) methods are inherently present in several scientific disciplines. As a subgroup of artificial intelligence, ML algorithms are capable of identifying particular patterns in data-sets by gaining experience (MURPHY, 2012; SAMMUT; WEBB, 2011; ALPAYDIN, 2014). The learning-ability of a machine is originally approached in Turing (1950). The development of ML techniques is strictly related to the data-processing capability, severely restricted until the end of nineties. As stated in Georgiev (2021), the advance of computational resources has dramatically improved the applicability of ML algorithms. These are applied in a broad variety of themes (SAMUEL, 1959; BISHOP, 2006; ABDULKADER; LAKSHMIRATAN; ZHANG, 2016; MURAKAMI; MIZUGUCHI, 2010; HORROCKS; HOLDEN; WEDGE, 2015; RAJKOMAR; DEAN; KOHANE, 2019; ORS, 2020; RAN et al., 2021). In Geosciences, ML methods are extremely required for solving some challenging problems. For example, Delfiner, Peyret e Serra (1987) produces a pioneer work using statistical analysis to predict lithologies from well-log data. An automatic lithologic description within formations are performed by combining wireline measurements with lithofacies into the method of discriminant analysis. Coudert, Frappa e Arias (1994) have presented a statistical approach using Bayesian discriminant analysis to predict lithofacies. A combination of several parameters represented by different Gaussian distributions are considered as prior information. Additionally, the prior probability is obtained from particular geological knowledge of the studied area. Following Bayes rule, the maximum posterior probability and the most useful log-responses to define lithofacies have been selected. Unfortunately, examples have shown that the success of prediction from wireline logs depends especially on the quality of the acquired data. More recently, Dell'Aversana, Ciurlo e Colombo (2018) integrate gravity, well-log, seismic and electromagnetic data using different ML methods to produce geologic models. Following the same guideline, Jia et al. (2021) uses an ensemble of ML methods combined with density, susceptibility and well-log data for the same purpose.

The obtained results reaches 90% accuracy. Bray e Link (2015) compares a support vector machine, a random forest and a neural network to identify UXO (Unexploded Ordnance) through magnetic data. In this case, the combination of methods provides 100% accuracy, with just 28% of false negatives. Yang et al. (2020) use a ML method to improve the resistivity modeling obtained by magnetotelluric inversion. (KUANG; YUAN; ZHANG, 2021) locate focal mechanism during an Earthquake using deep learning. Four events with amplitudes higher than 5.4 Mw are accurately determined. Wrona et al. (2018) test 20 supervised ML techniques to improve seismic interpretation by using previous interpreted data during the supervision stage.

Well logging concepts are based on acquisition tools in depth. The wide variety of equipment results in different geophysical logs. Consequently, various physical parameters are continuously measured in situ (ELKATTAN; ALFY; ELAWADI, 2020). This large set of observations allows a detailed study of the subsurface, which is prolific for exploration purposes (ELLIS; SINGER, 2007; DVORKIN; GUTIERREZ; GRANA, 2014). The affinity between rocks units and log readings is preponderant to the well-logging relevance to applied geophysics. However, there is no geophysical functional that associates log readings and the rock unit itself. Thus, there is plenty of room for ML methods in solving some challenging problems involving well-log data, such as simulation of logs (AKINNIKAWE; LYNE; ROBERTS, 2018; BADER; WU; FOMEL, 2018; HOSSAIN et al., 2020; FENG; GRANA; BALLING, 2021; MELLO; LUPINACCI, 2022) and facies identification (BUSCH; FORTNEY; BERRY, ; ROGERS et al., 1992; COUDERT; FRAPPA; ARIAS, 1994; BHATT; HELLE, 2002; SINGH, 2011; XIE et al., 2018; ALZUBAIDI et al., 2021; XU et al., 2021b). The former is a natural and well-established scientific challenge addressed to ML practitioners. This recognition problem is definitively a relevant issue in hydrocarbon exploration. Geophysical well-log data provide prolific vertical resolution and also good continuity among rock units (ELLIS; SINGER, 2007). Therefore, the classification of lithologies based on well-log data is one of the basis of reservoir studies. In the recent years, several researchers studied this problem. Li e Anderson-Sprecher (2006) identify lithofacies from well-log data by comparing both discriminant analysis and naïve Bayes classifier. The results indicate that both methods perform adequately for the given data set (i.e., GR, NPHI and resistivity logs). Li, Chan e Nguyen (2013) uses two specific ML methods to determine lithofacies log. A neural network and a decision tree are considered and the former provides the best classification outcome. More recently, Xu et al. (2021a) applied active learning to establish the lithology identification model with as few labels as possible. Popular ML algorithms are considered and extensive experiments to compare the performance of each algorithm are taken into account. With very few labeled data, the accuracy of the best method (i.e., Uncertainty-Entropy) reached $88,4$ %. (LAN et al., 2021) have proposed the multiclass positive and unlabeled machine learning (PU-learning) which uses labeled data and unlabeled data simultaneously to identify five types of carbonate log facies. This method can further use the log-data information of a

large number of unlabeled samples to alleviate the overfitting caused by insufficient labeled samples. Additionally, no additional labeling costs for model construction is considered. An actual application for two carbonate well blocks in the Tahe Oilfield shows that when using only limited labeled log facies data, PU-learning outperforms supervised support vector machine (Supervised-SVM) and supervised artificial neural network (Supervised-ANN) in general. Kumar, Seelam e Rao (2022) applied supervised ML techniques such as Support Vector Machine (SVM), Decision Tree (DT), Random Forest (RF), Multi-Layer Perceptron (MLP), and Extreme Gradient Boosting technique (XGBoost) for meticulously interpreting such banded coal seams from geophysical logs. To investigate the efficacy of the above-mentioned five ML techniques, Gamma-ray, Density, and Resistivity logs were considered from four boreholes drilled in Talcher coalfield, Eastern India. ML model training results indicate that all the prediction models have scored more than 88% accuracy scores in classifying carbonaceous and non-coal lithofacies.

Historically, the pioneer methods in facies classification are stochastic-based, which is a fundamental aspect for solving problems with inherent randomness. This kind of configuration is naturally encountered in well-log data. Thus, Naïve-Bayes (NB) classifiers emerges as one of the most prominent and effective inductive learning algorithms (RISH, 2001; ZHANG, 2005; QIN et al., 2017). Conceptually, a Bayesian approach to classification is appealing due to flexibility of new information to be introduced into prior probabilities and likelihood models (LANGLEY; IBA, 1992; TANG; WHITE, 2008). As stated in Domingos e Pazzani (1997), Rish (2001), the success of naïve Bayes classifiers is weak-dependent on a refined fit to a probability distribution. Rather, an optimal classifier is then obtained if the actual and estimated distributions agree on the most-probable class. The impact of this simplification is often surprisingly small and early experience with naïve Bayes suggests reasonable lithofacies reconstructions when compared to ML methods that are burden of lengthy training required, such as neural networks (LI; ANDERSON-SPRECHER, 2006; LI; CHAN; NGUYEN, 2013; MASOUDI et al., 2012; FENG; GRANA; BALLING, 2021). So, the conditional independence assumption might not be a major concern to achieve a competitive performance of NB classifier, as also mentioned in Li e Anderson-Sprecher (2006). This intriguing aspect calls our attention and is one of the major scientific justifications of this work. A deeper understanding of data patterns that affect the performance of naïve Bayes is still required. Additionally, improving the calculation of both prior probabilities and likelihood models are recommended if one wants an optimal performance of naïve Bayes classifiers (LI; ANDERSON-SPRECHER, 2006; XIE et al., 2018; HORROCKS; HOLDEN; WEDGE, 2015; XU et al., 2021a).

In this work, we promote a comparative evaluation of five alternative strategies to compute prior probabilities and/or likelihoods to be used in the naïve Bayes classifier. The problem to be solved consists of classifying rock units from well-log data of Massapê Field, in Recôncavo Basin, northeast Brazil. We begin this paper with an overall description of

Recôncavo Basin, especially the rock units comprising Massapê Field. We then describe and locate the log-data to be considered into training data-set. Gamma-ray, sonic and logarithm of resistivity logs of twelve wells are considered in our experiments. After that, true lithologic logs are computed by density and neutron-porosity logs, following Freire et al. (2020).

The first strategy, named as standard naïve Bayes, consists of computing likelihoods by means of Gaussian distributions and prior probabilities by analyzing the entire training data-set. In the second strategy, we then use the Gaussian kernel density estimation, named as KDE, to compute the likelihoods in place of conventional normal distributions, following the same framework in (LI; ANDERSON-SPRECHER, 2006; XIANG; YU; KANG, 2016). The third strategy considered in our investigation is the committee architecture, which trains a separate classifier on each well's data. Each classifier independently assigns a lithology label to the presented depth interval. These labels are tallied, and the highest voted lithology is used as the final label (HORROCKS; HOLDEN; WEDGE, 2015). The fourth strategy works directly into the estimation of prior probabilities in an iterative-optimization procedure, named as tuning (XIE et al., 2018; LOPES; JORGE, 2018). The fifth and last strategy, named as CRC, consists of integrating the stratigraphic information of Massapê Field to set up prior probabilities in different depth zones. As such, we considered this novel application of real interest, once the method is associated with previous geologic information of the studied area. With this, we highlight that naïve Bayes is able to improve classification outcomes. Through confusion Matrices and some statistics (i.e., precision, recall and fscore values) for the CRC strategy, we demonstrate and extend the applicability of naïve Bayes classifier to solving rock-units recognition. At last, we would like to reinforce the robustness and reliability of this learning algorithm, especially when reasonable log-data is considered.

7.2 Geologic settings: Recôncavo Basin

Recôncavo Basin is an onshore sedimentary basin located in Northeast Brazil, Bahia state. One of the pioneer exploratory scenarios in South America, Recôncavo is considered a mature sedimentary basin, which oilfields have been achieved the highest production rates. The sedimentary evolution of this basin comprises a transition from shallow-marine sedimentation, known as Pedrão Member to the Cazumba Member, a Paleozoic lakeside system (NETTO; OLIVEIRA, 1985; CARLOTTO, 2006; COURA, 2006; GORDON; DESTRO; HEILBRON, 2017).

The aborted rifting systems Recôncavo-Tucano-Jatobá are related to the early stages of Gondwana's breakup. The basic structure of Recôncavo sedimentary basin is a half-graben, limited by faults in East and West (see Figure 32). In an overall perspective, the depositional system date from Neo-Jurassic is basically composed of lacustrine, fluvial

and eolic depositions. The early-rifting lacustrine strata of Eo-cretaceous comprises the shales of both Candeias and Maracangalha formations, and the turbidite units (Pitangas and Caruaçu), which are basically transported by mass flows (COHEN, 1985; L.F., 2000; OLÍVIO et al., 2007).

During the medium stages of Eo-Cretaceous, the silting of Recôncavo Basin is major associated with deltaic and fluvial depositions of Pojuca and São Sebastião formations, respectively. The latter depositional system is the major occurrence of Neogene (BRUHN et al., 1994; FREIRE et al., 2020).

7.2.1 Massapê Field

The hydrocarbon exploration is still prolific in Massapê Field especially due to reservoir rocks of Maracangalha Formation - Caruaçu Member. Turbiditic sediments are the major occurrences in this formation (FREIRE et al., 2020). The turbidity fluxes are interbedded with lake deposits of shales and small amounts of marls (CAIXETA et al., 1994; FREIRE et al., 2020). During a study of electrofacies and also by high-resolution stratigraphy, Freire et al. (2020) defines sandstones, shales, siltstones and shaly sandstones as the main occurrences in Massapê Field. The latter two are herein coupled and referred to as slurries. Additionally, the Caruaçu member is separated by three depth-zones of turbiditic systems, CR-1, CR-2 and CR-3 (MUTTI; NORMARK, 1991; FREIRE et al., 2019). In this work, we make use of this prior geologic information to increase the classification accuracy of NB classifier.

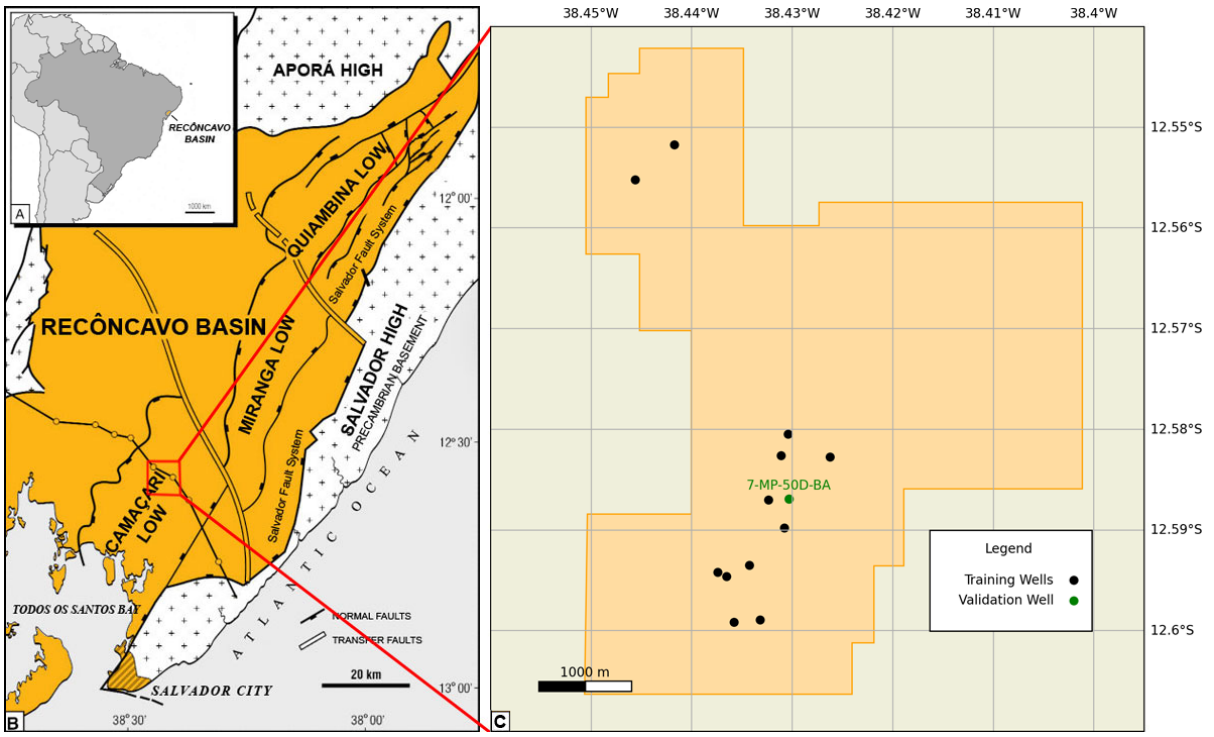


Figura 32 – (A) Location of the Recôncavo Basin in Brazil. (B) the study area comprising the Massapê oilfield. (C) The set of training wells (black dots) and the validation well (green dot) (BRUHN, 1999).

7.3 Well-log data of Massapê Field

All wells comprising the training database are composed by gamma-ray (GR), sonic (DT), logarithm of resistivity (log_ILD), bulk density (RHOB) and neutron porosity (NPHI) logs. The wells are located in Massapê oilfield, as can be seen in Figure 32 (b). There is a high number of drilled wells in the region, which increases the amount of reliable well-log data to be used in our training database. As a quality-control method, we discard all readings with compromised caliper information, resulting in a training dataset of twelve wells. We focus our application on Caruaçu Member within Maracangalha Formation (FREIRE et al., 2020).

7.3.1 Interpreted Lithologic Log: DRDN method

Commonly, the true lithologies are absence or not reliable in most of the wells, due to technical difficulties during the data-acquisition process. Alternatively, the interpretation of RHOB and NPHI logs is crucial for determining the so-called interpreted lithologic log, defined by the following expression:

$$\alpha_i = \frac{(\rho_{b_i} - 2.0)}{0.05} - \frac{(0.45 - \phi_{n_i})}{0.03}, \tag{7.1}$$

where i states for the i-th log value in depth, ρ_b and ϕ are RHOB and NPHI logs, respectively. The α_i is the i-th DRDN log-data used herein for interpreting the rock units in depth. For example, if $\alpha_i < -1.0$ then sandstone is assigned to the i-th depth. Shale is defined in case of $\alpha_i >= 0.3$. Finally, α values that lie between -1.0 and $+0.3$ states for slurry. As can be seen in Equation 7.1, DRDN log is a normalization factor involving ρ_b and ϕ_n logs. Such method is well suited for reservoir and non-reservoir rocks comprising the study area. To expand the DRDN method for other sedimentary basins, one should redefine the coefficients of Equation 7.1. For more details, readers are invited to (FREIRE et al., 2020).

7.4 Methodology

We start this section with some basic statistics to provide a qualitative error analysis. We then define the naïve Bayes (NB) classifier method in a standard way. After that, specific strategies to improve the standard (NB) application are considered. We hope that through those we could broaden the applicability of this robust Bayesian method for lithologic reconstruction problems. In our work, all log-data values are normalized to produce non-biased likelihood estimations.

7.4.1 Basic statistics

In a classification problem, an estimated model is inherently related to imprecision. This aspect can be computed by statistical validation metrics, such as confusion matrix, precision, recall and fscore (MCKINNEY, 2012; GRUS, 2015; BRUCE; BRUCE, 2019).

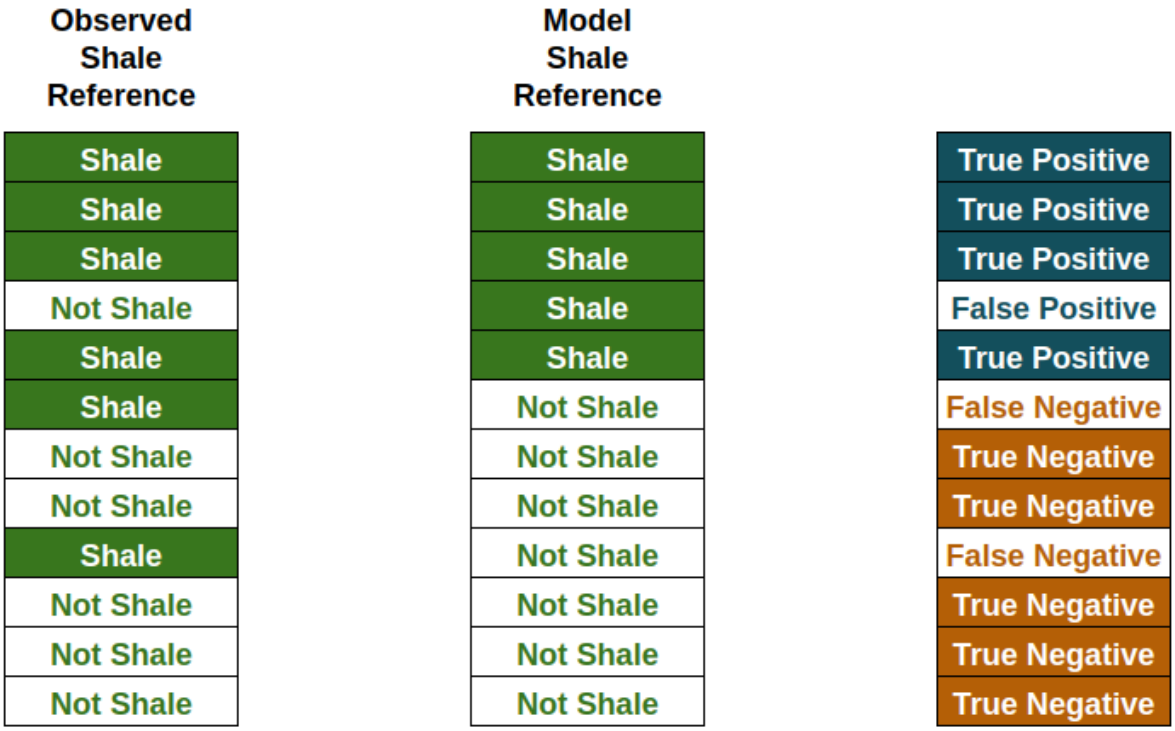


Figura 33 – Sketch of the concepts for true and false positives and negatives for a specific lithology (i.e., shale).

Once the true lithologic log is known, the predicted log can be quantified. Figure 33 illustrates the concepts of true positives (TP), true negatives (TN), false positives (FP) and false negatives (FN) for a fictitious lithologic log. Once these are clarified, we are able to define the confusion matrix (BRUCE; BRUCE, 2019) by combining TN and TF values into the principal diagonal, and the false quantities (FN and FP) off diagonal. Moving forward, we define precision as:

$$PR = \frac{\sum TP}{\sum TP + \sum FP}. \tag{7.2}$$

Equation 7.2 computes the true-positive rate of the positive elements presented into the confusion matrix. Recall is another useful quantity to be considered in an error analysis and is defined as:

$$RC = \frac{\sum TP}{\sum TP + \sum FN}. \tag{7.3}$$

In this case, Equation 7.3 is more sensitive to missclassifications of a particular lithology. Finally, the fscore value combines both precision and recall in the following:

$$FS = 2\frac{PR \times RC}{(PR + RC)}. \tag{7.4}$$

Equation 7.4 can be interpreted as the harmonic mean involving PC and RC values. This can be considered a more robust validation metric. As such, we use Equation 7.4

throughout the methodology section, once specific strategies are considered to improve NB classification outcomes.

7.4.2 Naïve-Bayes Classifier

Bayes theorem allows the prediction of a certain phenomenon through conditional probabilities, which is based on prior knowledge about the phenomenon and observations. A naïve Bayes (NB) classifier applies Bayes theorem where every feature is assumed to be class-conditionally independent (LI; ANDERSON-SPRECHER, 2006; MURTY; DEVI, 2011; LINDBERG; RIMSTAD; OMRE, 2015). This assumption simplifies the classification process by allowing the posterior probabilities to be calculated separately for each well log. Although the independence assumption is almost certainly violated, the NB classifier has been shown to be surprisingly robust in many domains that contain clear attribute dependences (LI; ANDERSON-SPRECHER, 2006).

In the context of this work, the likelihood is a conditional probability that considers a random log-data response for a particular lithology in depth. Based on this assumption, we can use the likelihoods and prior probabilities to classify lithologies. Considering all lithologies of your problem, the expected solution is the one with higher posterior probability. Through the naïve Bayes classification method, a particular lithology l at depth z is defined as l_z and can be obtained by:

$$l_z = max \left[\frac{p(l_i) \prod_{j=1}^{n} p(d_j|l_i)}{\sum_i p(l_i) \prod_{j=1}^{n} p(d_j|l_i)} \right], \tag{7.5}$$

where $p(l_i)$ is the prior probability of the $i-$th lithology and $p(d_j|l_i)$ represents the likelihood between the $j-$th log data (i.e., d_j) in depth and the $i-$th lithology (i.e., l_i) ((MURTY; DEVI, 2011)). Symbol max states the maximum computed posterior probability. Equation 7.5 is computed for each lithology in depth. The classification is then associated with the highest posterior-probability value. Commonly, priors are established by the knowledge of the entire set of lithologies and likelihoods are obtained by density-estimation functions.

7.4.3 Probability Density estimation

Probability density functions (PDFs) convert continuous data-sets to likelihoods. In this work, the well-log data is the amount of information to be represented by density functions. Two possibilities are considered in our approach: The conventional normal or Gaussian distribution and the Kernel Density Estimation (KDE) (ROSENBLATT, 1956; TARTER; KRONMAL, 1976; SILVERMAN; JONES, 1989). The former is commonly used in the standard NB classifier. The latter is extremely powerful for multi-modal data-sets.

7.4.3.1 Normal distribution

Normal probability functions are often assumed for data in practical situations. In this study, we assume log readings d given a certain facies (or lithology) l are normally distributed, with a probability density function given by:

$$N(d|l) = \frac{1}{\sqrt{2\pi\sigma_l^2}} \, e^{-\frac{1}{2\sigma_l^2}(d-\mu_l)^2}, \tag{7.6}$$

where σ_l^2 and μ_l are the variance and the mean of log readings given a lithology l, respectively. The normal distribution is required when a log data of a particular lithology is mono-modal. Additionally, the central limit theorem states that a large amount of data reasonably fits a normal distribution (CLAPHAM; NICHOLSON, 2009; UPTON; COOK, 2014).

7.4.3.2 Kernel Density Estimation

In this particular case, the density function is approximated by the superposition of a set of kernels (LI; ANDERSON-SPRECHER, 2006). As a non-parametric distribution, the Kernel Density Estimation (KDE) is entirely related to the data to be fitted by a continuous distribution per lithology l:

$$K(d|l) = \frac{1}{nh} \sum_{i=1}^{n} N\left(\frac{d-d_i}{h} \mid l\right), \tag{7.7}$$

where n is the log-data sampling, h is the window width computed by (SCOTT, 2015), N is normally a symmetric probability density function, the normal distribution in our case. The training log data and the $i-$th log data are d and d_i, respectively. KDEs are particularly recommended for multi-modal likelihoods (ROSENBLATT, 1956; TARTER; KRONMAL, 1976). Additionally, this nonparametric density function is more sensitive to data variability then parametric ones (SILVERMAN, 1986; SILVERMAN; JONES, 1989; BIANCHI, 1995).

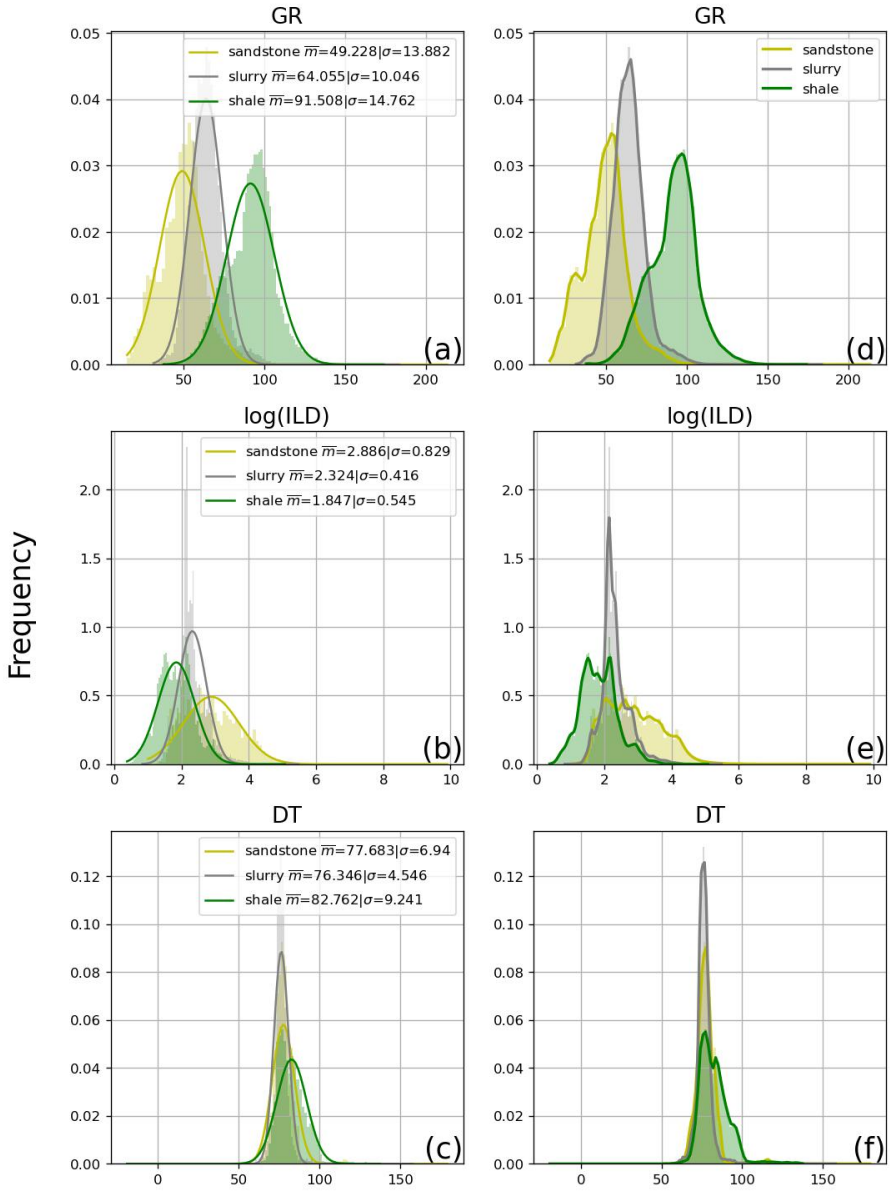


Figura 34 – Likelihoods computed by normal distribution (a, b, c) and using the Gaussian kernel density estimation (d, e, f) for each lithologies of the entire training data-set.

Figures 34 (a), (b) and (c) show likelihoods obtained by means of the normal distribution (i.e., Equation 7.6). Means and standard deviations for sandstones, shales and slurries are also computed for each log-data, totaling an amount of nine parameters. Additionally, we also present the likelihoods when the Gaussian KDE is considered, as can be seen in Figura 34 (d), (e) and (f). We clearly observe the improvements in data-representation that KDE can provide, especially for multi-modal data-sets. In counterpart, the KDE application in a high-definition bandwidth is computationally severe and unusual. In this work, we use the algorithm implemented in Virtanen et al. (2020), which is based on Scott (2015) for computing the optimal bandwidth.

7.4.4 Standard naïve Bayes (NB) classifier

We define the standard NB classifier as the one using likelihoods computed by Equation 7.6 (Figura 34). Prior probabilities are calculated by the knowledge of proportions of sandstones, slurries and shales in the entire training data-set, following Li e Anderson-Sprecher (2006), Murty e Devi (2011). In our context, the lithologies are defined by the DRDN method (Equação 7.1).

In this work, we make use of different strategies to compute the likelihoods and/or priors (i.e., elements of Equation 7.5) to improve the NB classifier outcomes. From herein, the standard NB classifier is referred to as STD.

7.4.5 Strategy 1: likelihoods using the Gaussian Kernel Density Estimation

In this first case, we use the Gaussian KDE (i.e., Equation 4.11) to set the likelihood models instead of using the standard normal distribution (i.e., Equation 7.6). This particular strategy is referred to as KDE. Figure 34 shows both normal and Gaussian KDE likelihoods for the training data-set comprised by wells of Massapê oilfield. From the DRDN perspective, we can clearly observe a separation between shales and sandstones, with slurries in-between. This overall data-configuration indicates a promising NB classification outcome.

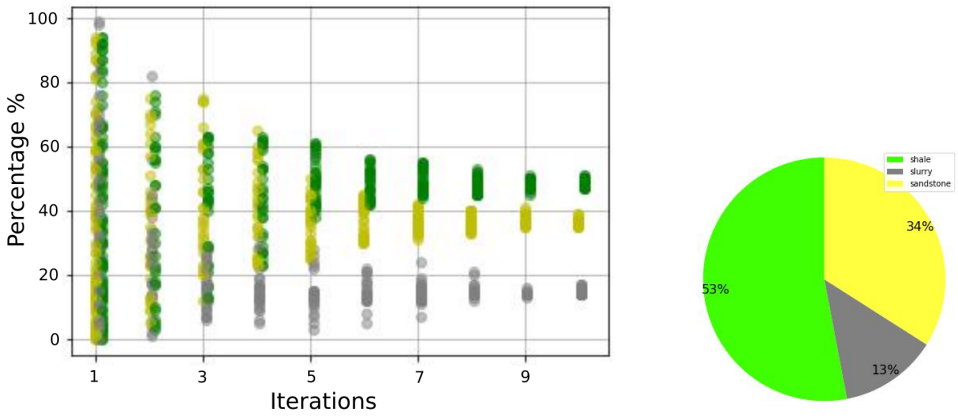
7.4.6 Strategy 2: Automatic definition of priors - Tuning

A challenging aspect of the NB classifier lies in a promising definition of prior probabilities (i.e., $p(l_i)$ in Equation 7.5). Considering the prior probabilities as hyperparameteres to be optimized, we define a tuning (XIE et al., 2018; SIDDIG et al., 2021) process by the following steps:

1. Definition of N priors, randomly selected within a search limits of $[0, 100\%]$ [1] per lithology;
2. Use the cross-validation method (XIE et al., 2018) to evaluate the classification performance of the STD strategy for the set of N priors;
3. Select a subset of priors based on the highest fscore sum values (i.e., Equation 7.4);
4. Reset the search limits for a minimum and a maximum prior values for each lithology based on the previous subset;
5. Repeat the process narrowing the search limits;
6. Reach the best prior after M iterations.

[1] the sum of each prior must be equal to 100.

The above-mentioned iterative procedure guarantees the convergence to a final prior-probability model, as can be seen in Figures 35 (a and b). From herein, the tuned NB classifier is referred to as TUN.



(a) (b)

Figura 35 – (a) Convergence of priors using M=10 iterations and (b) the best obtained prior.

7.4.7 Strategy 3: Altering Training data-sets, priors and likelihoods - Architecture

The architecture process explores different training data-set configurations to produce a robust classification. In our context, we implement the committee architecture, which is related to a single classification model obtained from independent training sets (i.e., composed by single wells) (HORROCKS; HOLDEN; WEDGE, 2015). Basically, we can summary the procedure by the following steps:

1. Definition of M individual training data-sets composed by single wells;
2. Compute M likelihoods and priors (i.e., $p(d_j|l_i)$ and $p(l_i)$ in Equation 7.5, respectively) for each individual training data-set;
3. Apply M standard NB classifier into the validation well;
4. Find into the M classified models the most recurrent lithology for each depth.

The committee architecture individually covers the entire training data-set and explores the generality of each obtained classification model.

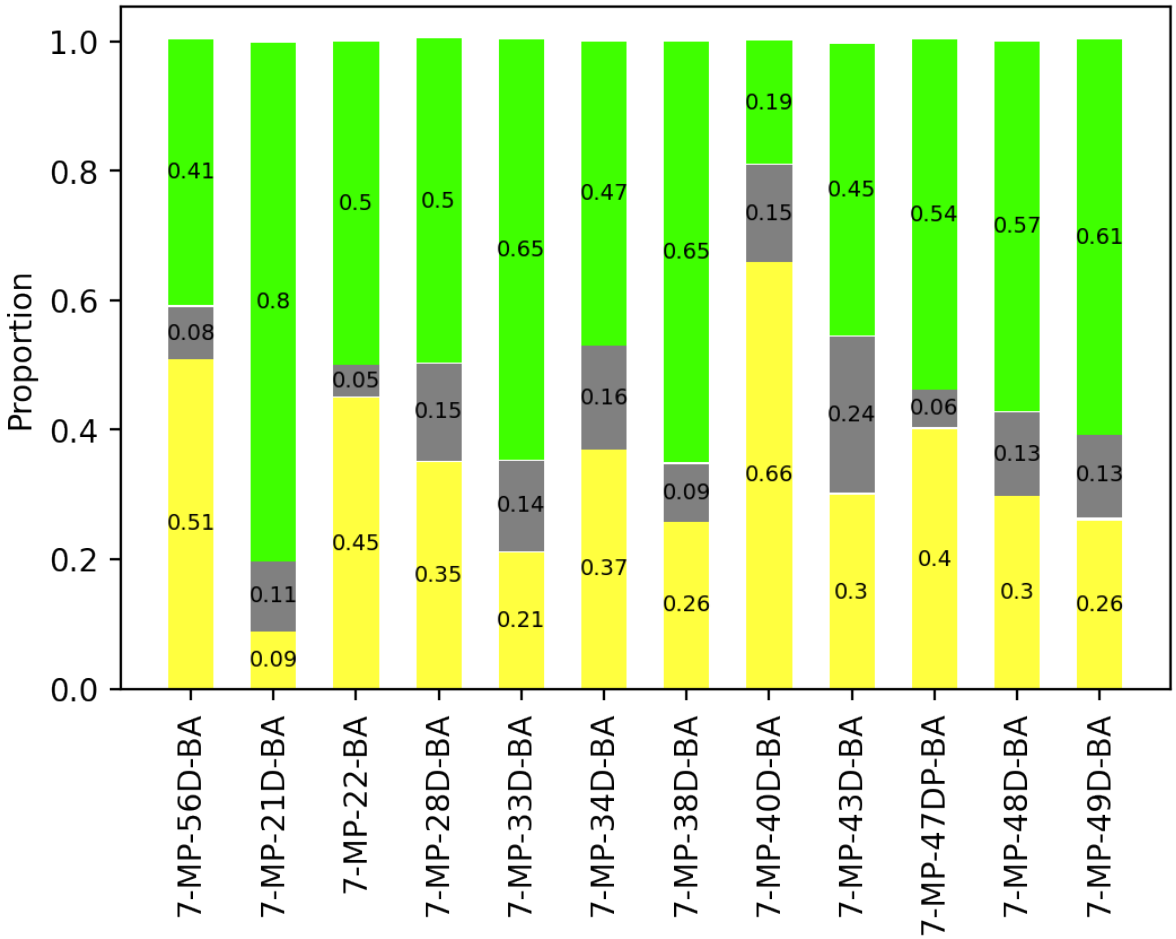
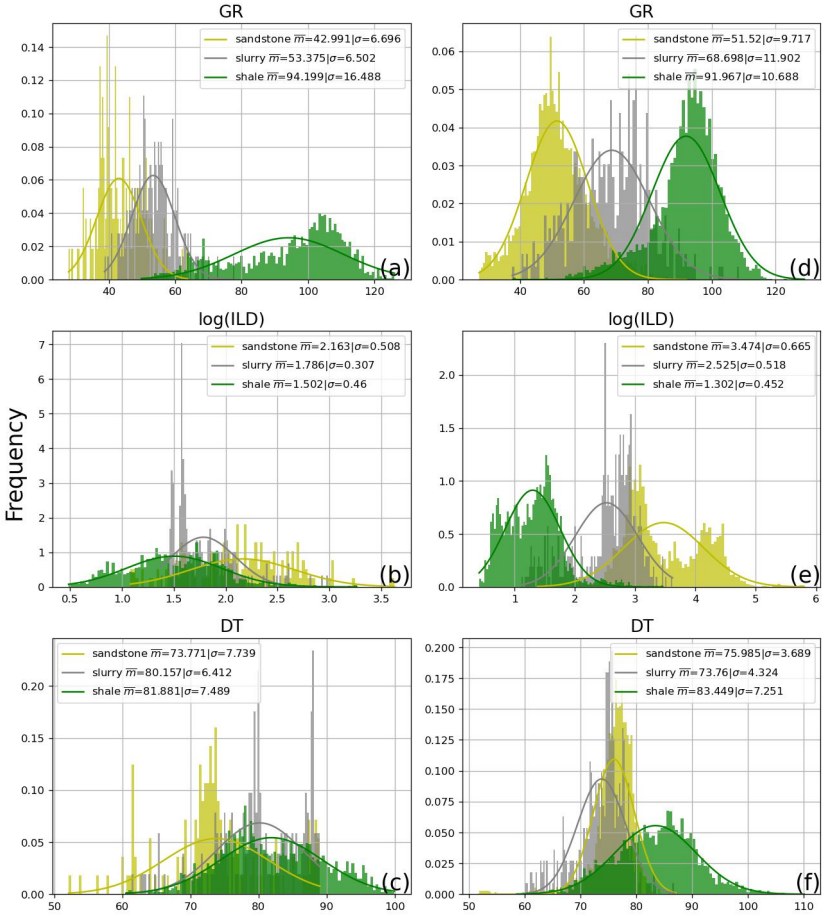


Figura 36 – Bar graph with all priors obtained for each training well.

Due to the large amount of data, plotting all likelihoods of each well is completely unfeasible. As an alternative, we present two extreme situations concerning the likelihoods of individual wells. Figure 36 shows the likelihoods computed for each well of the individual training data-set, as mentioned in item 2.



(a) (b)

Figura 37 – The worst (a, b and c) and the best (d, e and f) likelihood models obtained during the architecture procedure. $\bar{m}$ and σ are means and standard deviations for sandstones (yellow), shales (green) and slurries (grey).

Figures 37 (a,b,c) and (d,e,f) show the worst and the best set of likelihoods comprising sandstones, shales and slurries, respectively. We use the fscore value (i.e., Equation 7.4) to rank the outcomes obtained during the computation of the committee architecture. From herein, the combination of standard NB classifier (STD) and the committee architecture procedure is referred to as ARC.

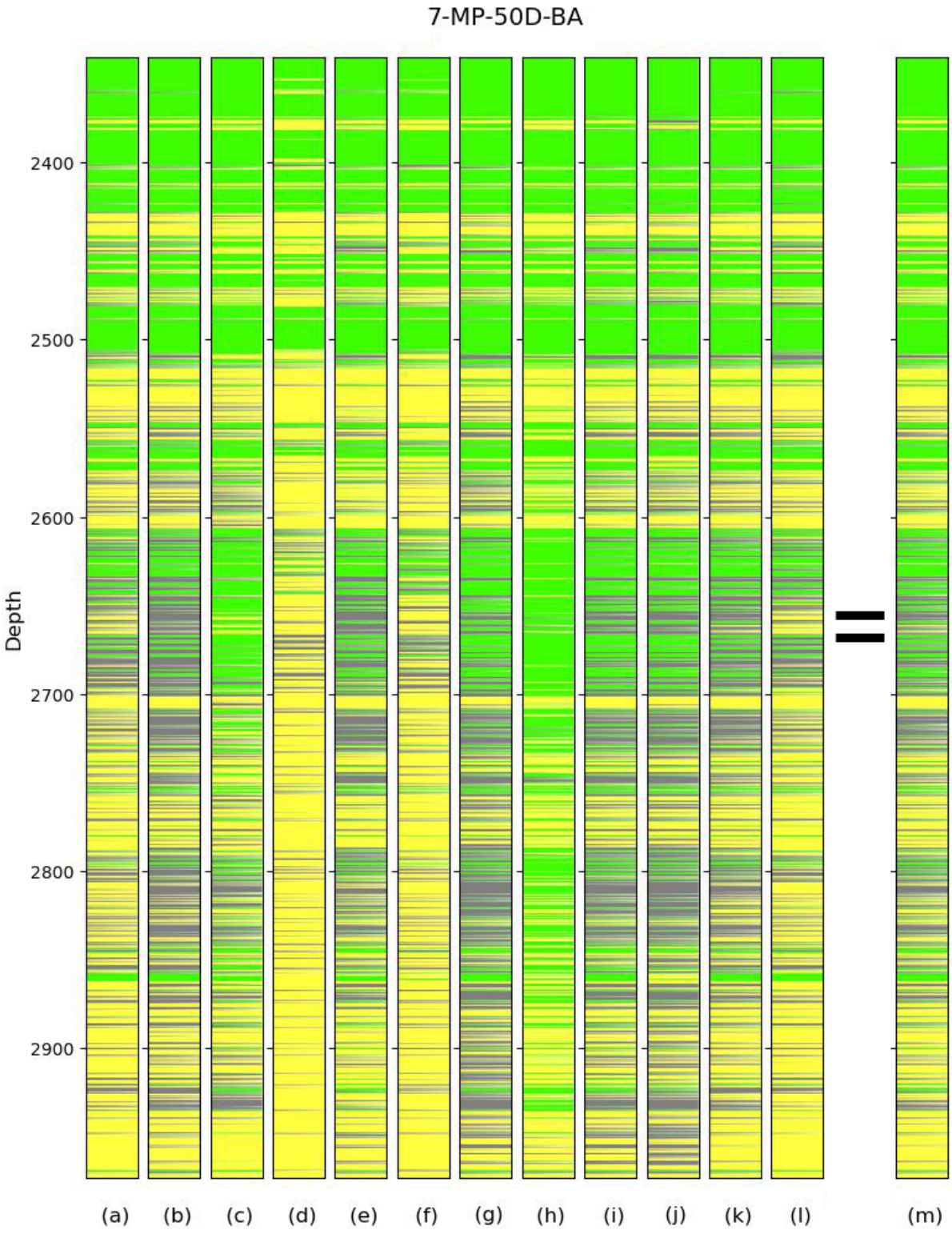


Figura 38 – All classifications of the validation well 7-MP-50D-BA (a-l) obtained during the architecture procedure. Track m shows the final classification outcome.

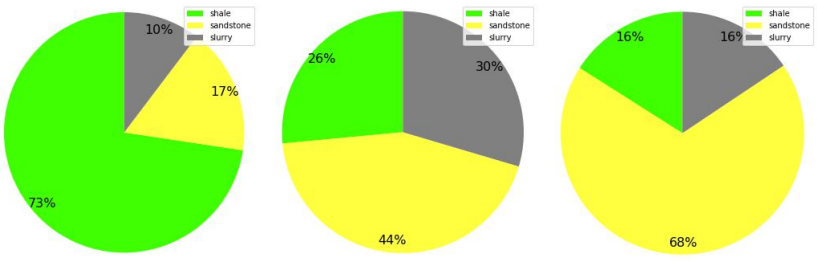
As a measure of the processing time during the ARC strategy, we also present the classifications obtained by means of individual training data-sets, as can be seen in Figures 38 (a-l). So, the final classification outcome (i.e., Figure 38 m) is the most recurrent

lithology for each depth.

7.4.8 Strategy 4: Using stratigraphic information in depth-zones

In this case, we adapt the standard NB classifier to be applied in particular depth-zones within the validation well. For each zone, a different prior probability is considered. The lithologic proportions are then based on the production depth-zones of Caruaçu member, in Maracangalha formation, as mentioned in Freire et al. (2020). The valuable aspect of this application lies in the use of stratigraphic information to split the standard NB classification. As such, we name this strategy as CRC. The training data-set is the same used in standard NB classifier (i.e., likelihoods computed by Equation 7.6).

Figure 39 (a, b and c) show three different prior probabilities used in CRC application. The CRC-1 depth-zone is majority covered by shales, while CRC-2 and CRC-3 present more sandstones (44% and 68%, respectively). We define all depth zones according to Freire et al. (2020).



(a) First zone (b) Second zone (c) Third zone

Figura 39 – Pie plots showing different prior probabilities based on depth-zones for validation-well 7-MP-50D-BA.

Figura 40 presents a summary of the entire methodology considered in this work.

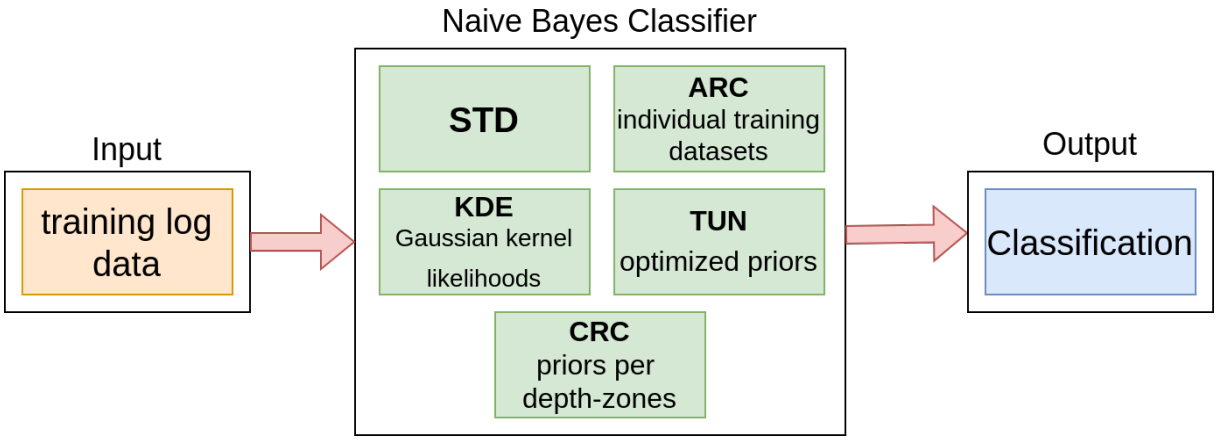


Figura 40 – Flowchart of the proposed method.

7.5 Results

In this section we present the classification of lithologies obtained by standard NB and all strategies presented in this work. The well-log data consist of gamma-ray (GR), sonic (DT), and deep resistivity (ILD). Formation density (RHOB) and neutron porosity (NPHI) logs are used to compute the interpreted lithologic log by DRDN method (i.e., Equation 7.1). The resistivity data are log normally distributed and designated by log(ILD). Among the three well logs, the variables NPHI, DT, and log(ILD) show moderately overlapping with each other, as shown in Figura 34 and Figura 37. Specifically, the log(ILD) log presented in Figura 37 (e) shows a bimodal pattern for sandstones, which might be related to the presence of hydrocarbons.

We then define the most promising result by analyzing the maximum mean value of fscore (Equation 7.4) for each well of the data-set. After that, we show a detailed view of the results for the validation well.

7.5.1 Selecting the validation well

We applied all strategies in all wells of Massapê oilfield to verify the most promising outcomes. Figure 41 presents fscore values for standard NB classifier (STD) and all strategies (i.e., KDE, ARC, TUN and CRC). The choice for the validation well comprises the most promising strategy and is achieved by computing the highest fscore mean value for the CRC strategy. Additionally, the larger fscore discrepancies per well are also considered as a selection criterion. So, we select 7-MP-50D-BA (CRC = 0.76, STD = 0.72, KDE = 0.71, ARC = 0.71 and TUN = 0.73) as the validation well.

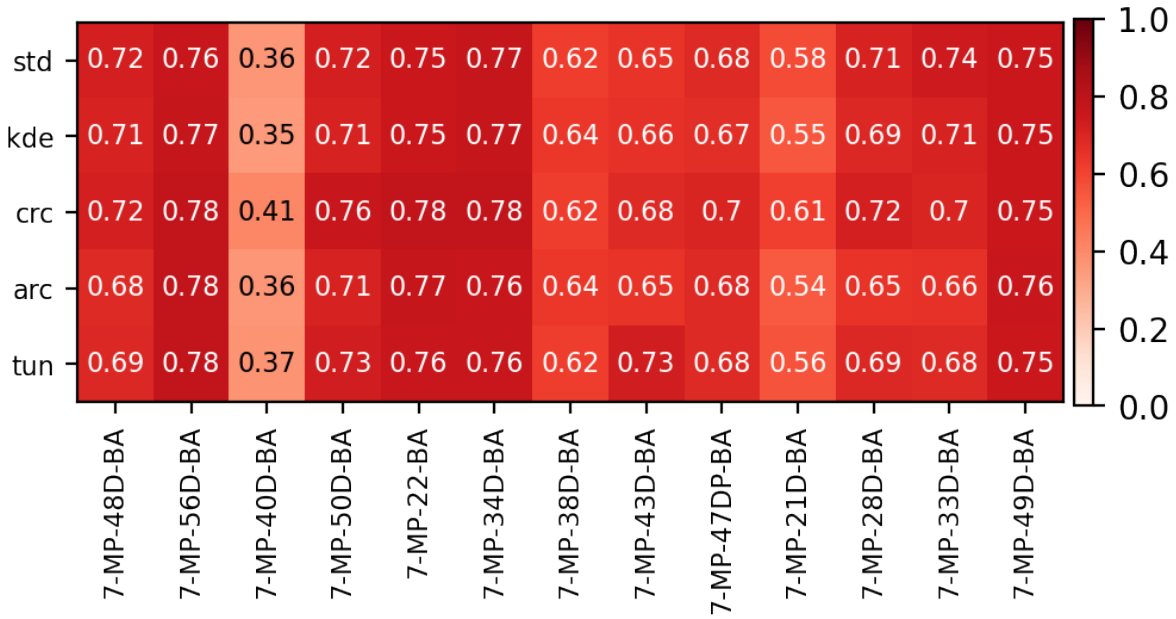


Figura 41 – Error analysis by using the fscore calculation (i.e., Equation 7.4).

7.5.2 Validation well: 7MP-50D-BA

Figure 42 shows the classifications for the validation well obtained by each strategy. From a general perspective, all NB classifier strategies produced similar reconstructions, which reinforces the robustness of such Bayesian method for classification of lithologies.

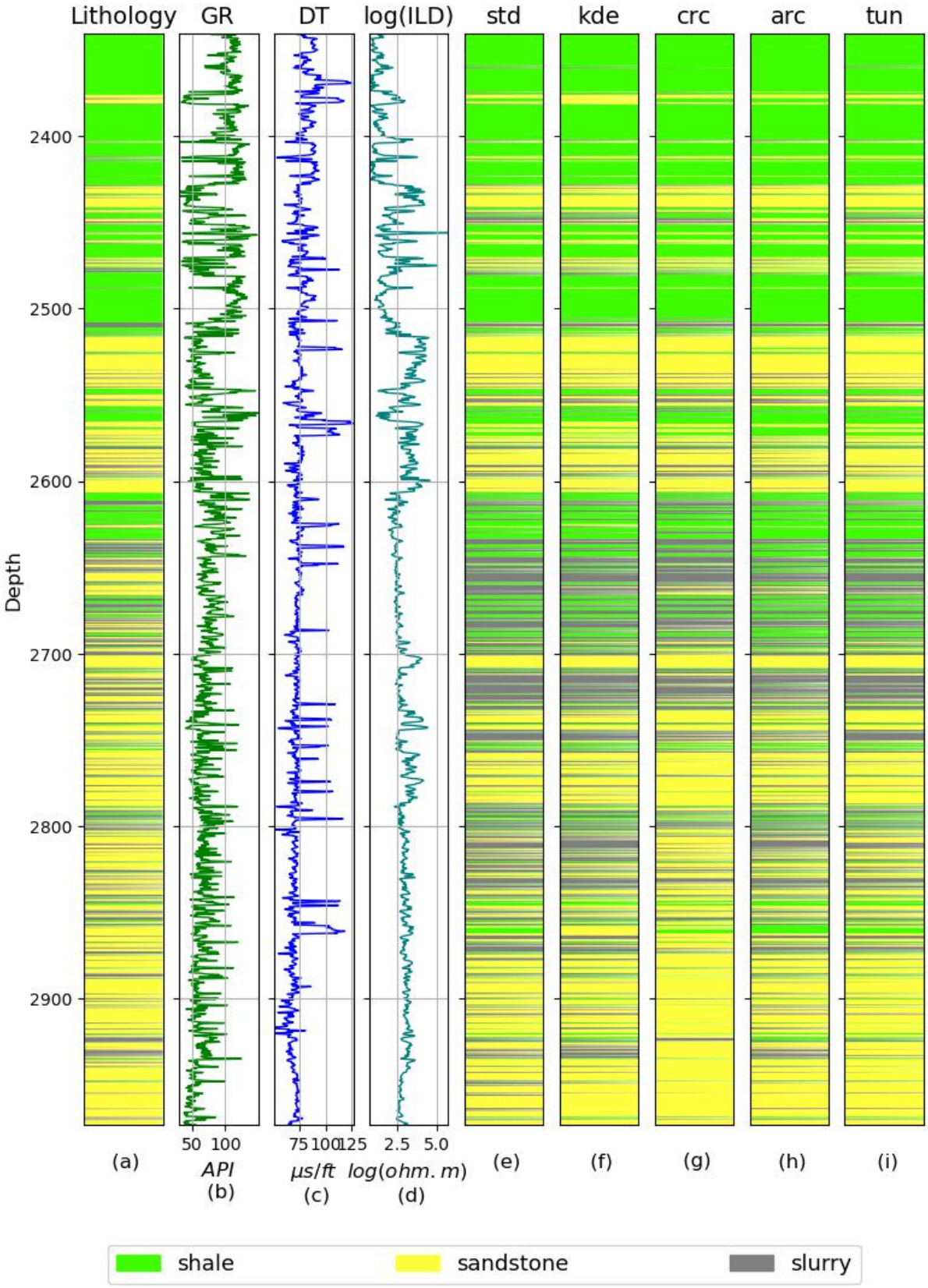


Figura 42 – Overall classifications obtained for 7MP-50D-BA. (a) Interpreted Lithologic Log, (b) GR log, (c)DT log, (d) log(ILD) log. The classification outcome using (e) standard NB, (f) KDE, (g) CRC, (h) ARC, (i) TUN.

Particularly for depth-ranges between 2600 and 2700 m, two thin packages of sandstones presented in Figure 42(a) are miss-classified by all strategies. This pattern might be related to the lack of information into the resistivity log (Figure 42-d). Additionally, the absence of hydrocarbons in reservoir rocks at those depths is also noticeable. We can observe that the CRC strategy is unduly the best for classifying reservoir rocks bellow 2800 m depth. In opposition, strategies STD, KDE, TUN and ARC are induced for more slurries than observed in the true lithologic log, as shown in Figures 42(e-i). So, we can affirm that the CRC strategy is well suited for classification of lithologies of Massapê oilfield, especially when GR, DT and log(ILD) logs are considered. In the next section, we present a more in deep analysis of the CRC strategy applied to the naïve Bayes classifier.

7.5.3 Classification of well 7MP-50D-BA using the CRC strategy

Figures 43 (a-h) present the results for validation well 7MP-50D-BA when naïve Bayes classifier uses stratigraphic information. Figure 43 (a), (b), (c) and (d) show the interpreted lithologic log obtained by DRDN method, GR, DT and log(ILD) logs, respectively. Improving the classification analysis, we also present an error log in Figure 43 (f). In an overall perspective, misclassifications are more often at the CRC-2 and CRC-3 zones for slurries and sandstones (i.e., turbidites). As expected, the determination of shales are more accurate in all zones due to well-defined likelihoods, as previously mentioned.

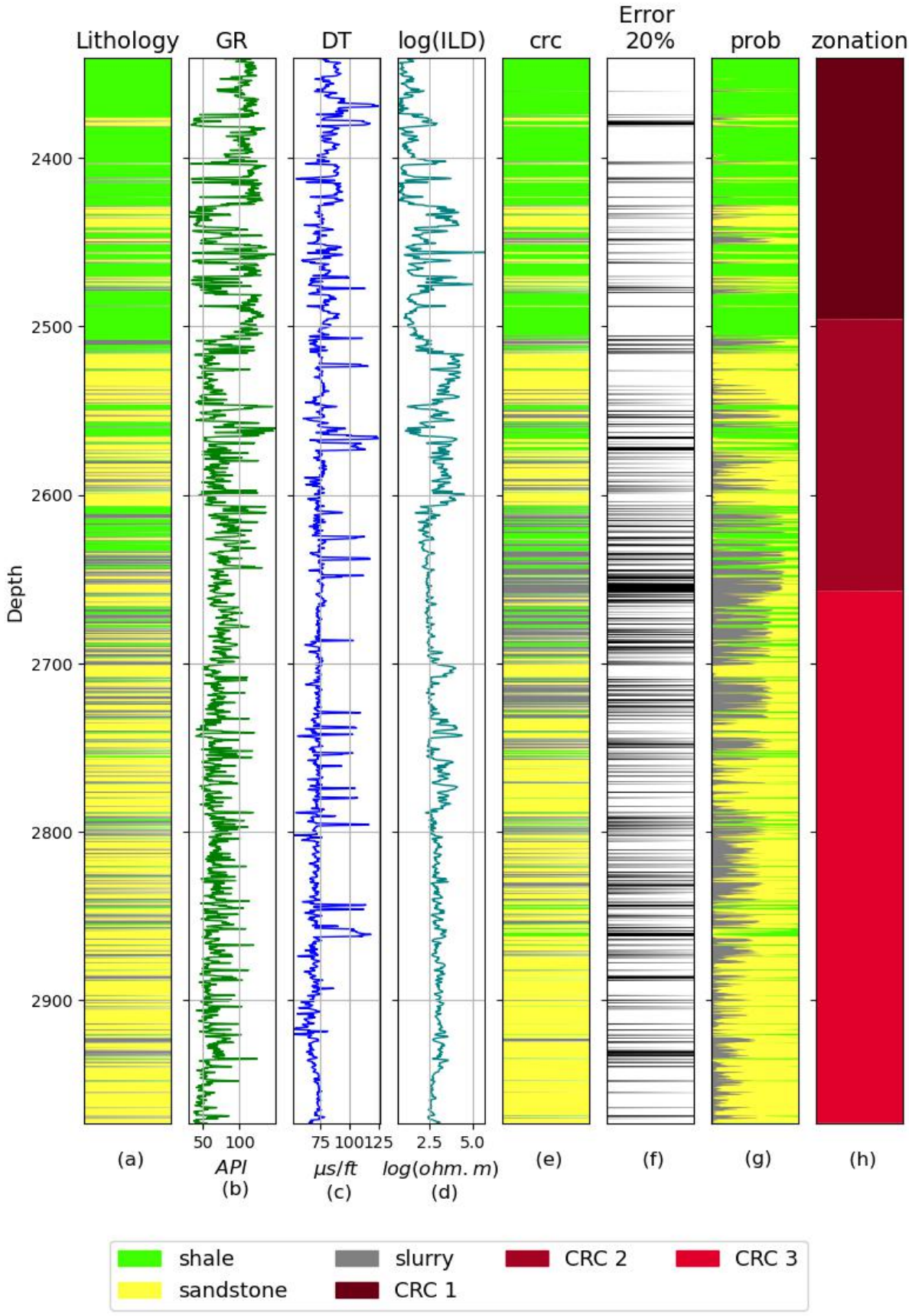
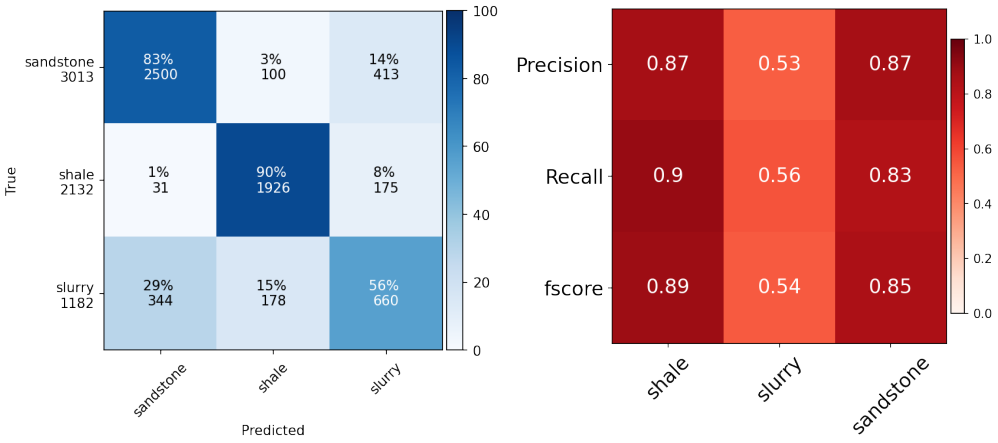


Figura 43 – Results for validation well 7MP-50D-BA. (a) Interpreted Lithologic Log, (b) GR log, (c) DT log, (d) log(ILD) log, (e) classification log, (f) error log, (g) probability log, (h) depth-zones. The zone CRC-1 varies from 2340 to 2495 m depth, while zone CRC-2 varies from 2495 to 2657. The CRC-3 zone goes from 2657 down to 2973 m depth.

Figure 43 (e) exhibits the classified log by means of CRC strategy. We can clearly observe that the definition of prior probabilities for each depth-zone reduces the total amount of errors. As an example, the DRDN log (Figure 12a) shows large packages of sandstones below 2650 m depth, which is indeed reasonably represented by the classified log. An error log (i.e., Figure 43 f) is also considered, showing a reasonable error of 20%. Major errors are encountered at zone CRC-2 (i.e., around 2650 m depth). This aspect might be related to the increased amount of slurries at those depths. Additionally, the geologic settings for the validation well present less hydrocarbons at the transition between CRC-2 and CRC-3 zones.



(a) (b)

Figura 44 – (a) Confusion Matrix and (b) Precision, recall and fscore of CRC (i.e., Equations 7.2, 7.3 and 7.4, respectively) strategy applied to the validation well 7MP-50D-BA.

A more in deep statistical analysis of errors is provided by confusion matrix, as can be seen in Figure 44 (a). In an overall perspective, shales are more efficiently classified than sandstones and slurries. Additionally, slurries and sandstones are more often confounded than slurries and shales. Figure 44 (b) exhibits the precision, recall and the fscore values for the validation well. We can observe identical precision of CRC in classifying shales and sandstones. This aspect is due the larger amount of sandstones in relation to shales. In counterpart, slurry presents the worse precision.

7.6 Discussions and Conclusions

An accurate naïve Bayes application depends on promising prior probabilities and likelihoods. Estimation of probability densities is important for the calculation of the likelihood and thus for estimation of lithologies. Prior probabilities are challenging to be set due to its inherent subjectivity. To cover such particularities, we present in this paper four strategies that illustrate the eventual improvements that Naïve Bayes (NB)

classifier can produce compared to the standard (NB) application. This latter method computes likelihoods by means of normal distributions and prior probabilities are obtained by exploring the lithologic information presented in the entire training data-set.

The first strategy, named as KDE, computes likelihoods by non-parametric density functions instead of normal distributions. This approach is expected to be promising when multi-modal data should be classified. For the validation well presented in this work, the comparison of KDE and the normal distribution indicates a similar classification performance, with a narrow vantage for the standard NB, following the same conclusions of Li e Anderson-Sprecher (2006). The speculation over the bandwidth is absolutely relevant for improvements in KDE classifications. Furthermore, compared with other density estimation methods, fitting log readings to a KDE is computationally demanding for a training data set basically composed by smooth and monomodal distributions.

The second strategy, named as TUN, represents prior probabilities by hyperparameters to be determined by some optimization strategy. In our case, a sort of Monte Carlo optimization method is implemented and a promising prior is automatically encountered by maximizing the fscore values. This might be interesting when information about the distribution of lithologies in the studied area is lacking. So, TUN works as an automatic estimator of priors and is recommended when lithologic interpretation is dubious, which is an aspect not considered in the standard NB classifier.

The third strategy, referred to as ARC, combines multiple versions of unstable classifications in a final and stable outcome compared to the standard NB classifier. In our work, we create several classification outcomes by means of single-well training data-sets. With this, we explore the variability of the training data-sets to produce more stable and robust classifications when compared to the standard NB classifier.

The fourth strategy, named as CRC, uses geologic information in depth to produce stratigraphic zones. This particularity is directly applied in standard NB classifier to produce more specific prior probabilities during the classification procedure. As such, relevant improvements comprising reservoir rocks at the bottom of the well are achieved when the validation well is splitted into three specific stratigraphic zones. The CRC strategy can be considered the major innovative aspect of this work, once the combination of geologic information and a Machine Learning method can significantly improve reservoir analysis by means of log data. Additionally, we encourage practitioners to expand such application to other stratigraphic zones, apart from the exploration of hydrocarbons.

The above-mentioned strategies are applied into the lithologic classification problem comprising well-log data of Massapê oilfield, in Recôncavo Basin, Northeast Brazil. As a quality-control method, basic statistics, such as confusion matrix, precision, recall and fscore are also considered.

Commonly, the naïve Bayes (NB) classifier is based on the principle of independence of features. This assumption is considered one of the most contradictory aspects of NB classifiers (CRACKNELL; READING, 2014; HORROCKS; HOLDEN; WEDGE, 2015; XIE et al., 2018; CHERAGHI; KORD; MASHAYEKHIZADEH, 2021). As example, Li e Anderson-Sprecher (2006) indicate that the independence assumption is not a major concern in their classification problem due to good obtained classifications. (MURAKAMI; MIZUGUCHI, 2010) is not assertive about the degree of independence among features in their biological problem and hence an application of NB classifier is worth attempting. Following a different outline, (DOMINGOS; PAZZANI, 1997) and (FRANK et al., 2000) verified that the naive Bayes (NB) produces undesired results when it is used for data-regression. In this work, we endorse the discoveries of (LI; ANDERSON-SPRECHER, 2006) for the classification problem involving well-log data. In conclusion, further investigations about this issue should be performed to broader the applicability of NB classifiers. An alternative lies in applying the augmented naïve Bayes classifier (ZHANG, 2005) to reconstruction of lithologies. In our application itself, the above-mentioned principle is definitively not an obstacle for reliable classifications, especially when the most probable lithology is associated to the highest likelihoods.

In this study, the standard and other four strategies associated to the naïve Bayes classifier are compared in the problem of classification of lithologies of Massapê oilfield, in Recôncavo basin, Northeast Brazil. In an overall view, we obtained a slightly improved classification when the CRC strategy is compared to the others. This is probably related to the stratigraphic evidences that are incorporated to the set of prior probabilities. We hope that through this work we have demonstrated the great advantages that NB classifiers associated with both geologic settings and alternative strategies can bring to machine learning applied to lithologic classification.

7.7 Acknowledgments

First author fully thanks CAPES Foundation, Ministry of Education of Brazil and Petrobras for the scholarship. We also thank ms. Carolina Ferreira da Silva for providing precious stratigraphic information to be used in one of the strategies presented in this work.

Referências

ABDULKADER, A.; LAKSHMIRATAN, A.; ZHANG, J. *Introducing DeepText: Facebook's text understanding engine - Facebook Engineering*. 2016. Disponível em: <https://engineering.fb.com/2016/06/01/core-data/introducing-deeptext-facebook-s-text-understanding-engine/>.

AGRAWAL, T. Hyperparameter optimization in machine learning. *Hyperparameter Optimization in Machine Learning*, Apress, 2021.

AKINNIKAWE, O.; LYNE, S.; ROBERTS, J. Synthetic well log generation using machine learning techniques. In: *SPE/AAPG/SEG Unconventional Resources Technology Conference 2018, URTC 2018*. [S.l.]: Unconventional Resources Technology Conference (URTEC), 2018.

AKKAŞ, E. et al. Application of Decision Tree Algorithm for classification and identification of natural minerals using SEM-EDS. *Computers and Geosciences*, Elsevier Ltd, v. 80, p. 38–48, jul 2015. ISSN 00983004.

ALPAYDIN, E. *Introduction to Machine Learning*. [S.l.]: MIT Press, 2014. (Adaptive Computation and Machine Learning series). ISBN 9780262028189.

ALZUBAIDI, F. et al. Automated lithology classification from drill core images using convolutional neural networks. *Journal of Petroleum Science and Engineering*, v. 197, p. 107933, 2021. ISSN 0920-4105. Disponível em: <https://www.sciencedirect.com/science/article/pii/S0920410520309888>.

ARTHUR, D.; VASSILVITSKII, S. K-means++: The advantages of careful seeding. In: . [S.l.: s.n.], 2007. v. 8, p. 1027–1035.

ASHARI, A.; PARYUDI, I.; TJOA, A. M. Performance comparison between naïve bayes, decision tree and k-nearest neighbor in searching alternative design in an energy simulation tool. *International Journal of Advanced Computer Science and Applications*, v. 4, 12 2013. Disponível em: <https://thesai.org/Downloads/Volume4No11/Paper_5-Performance_Comparison_between_Na%C3%AFve_Bayes.pdf>.

ASQUITH, G. B.; GIBOSN, C. R. Basic Well Log Analysis for Geologists. 1982.

BADER, S.; WU, X.; FOMEL, S. Missing well log estimation by multiple well-log correlation. In: EUROPEAN ASSOCIATION OF GEOSCIENTISTS & ENGINEERS. *80th EAGE Conference and Exhibition 2018*. [S.l.], 2018. v. 2018, n. 1, p. 1–5.

BHATT, A.; HELLE, H. Determination of facies from well logs using modular neural networks. *Petroleum Geoscience*, v. 8, p. 217–228, 09 2002. Disponível em: <https://www.researchgate.net/publication/275249542_Determination_of_facies_from_well_logs_using_modular_neural_networks>.

BIANCHI, M. Bandwidth selection in density estimation. Springer, New York, NY, p. 101–112, 1995. Disponível em: <https://link.springer.com/chapter/10.1007/978-1-4612-4214-7_6>.

BISHOP, C. M. *Pattern recognition and machine learning*. Springer, 2006. 738 p. ISBN 1493938436. Disponível em: <https://www.microsoft.com/en-us/research/uploads/prod/2006/01/Bishop-Pattern-Recognition-and-Machine-Learning-2006.pdf>.

BRAY, M. P.; LINK, C. A. Learning machine identification of ferromagnetic UXO using magnetometry. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, Institute of Electrical and Electronics Engineers, v. 8, n. 2, p. 835–844, feb 2015. ISSN 21511535.

BREIMAN, L.; BREIMAN, L. Bias, variance, and arcing classifiers. 1996. Disponível em: <http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.30.8572>.

BRUCE, A.; BRUCE, P. *Estatística Prática para Cientistas de Dados*. 1. ed. [S.l.]: Alta Books, 2019. ISBN 9788550810805.

BRUHN, C. H. L. Reservoir architecture of deep-lacustrine sandstones from the early cretaceous recôncavo rift basin, brazil. *AAPG Bulletin (American Association of Petroleum Geologists); (United States)*, v. 83, p. 1502–1525, 1999.

BRUHN, C. H. L. et al. Reconcavo basin, brazil: A prolific intracontinental rift basin: Chapter 5: Part ii. examples of other rift basins. AAPG Special Volumes, v. 137, p. 157–203, 1994.

BUSCH, J. M.; FORTNEY, W. G.; BERRY, L. N. Determination of lithology from well logs by statistical analysis. *SPE (Society of Petroleum Engineers) Format. Eval.; (United States)*, v. 2. Disponível em: <https://www.osti.gov/biblio/5537199>.

CAIXETA, J. et al. *Bacias do Recôncavo, Tucano e Jatobá*. Rio de Janeiro: [s.n.], 1994. 9 p.

CARLOTTO, M. A. *ANÁLISE ESTRATIGRÁFICA DOS FLUXOS GRAVITACIONAIS DA FORMAÇÃO MARACANGALHA NO CAMPO DE JACUÍPE, BACIA DO RECÔNCAVO, BRASIL*. Tese (Dissertação) — UNIVERSIDADE FEDERAL DO RIO GRANDE DO SUL (UFRGS), Porto Alegre, dec 2006.

CARREIRA, V.; PONTE-NETO, C.; BIJANI, R. A comparison of machine learning processes for classification of rock units using well log data. In: EUROPEAN ASSOCIATION OF GEOSCIENTISTS & ENGINEERS. *80th EAGE Conference and Exhibition 2018*. [S.l.], 2018. v. 2018, n. 1, p. 1–5.

CASELLA, G.; BERGER, R. *Inferência Estatística*. [S.l.]: CENGAGE DO BRASIL, 2010. ISBN 9788522108947.

CHERAGHI, Y.; KORD, S.; MASHAYEKHIZADEH, V. Application of machine learning techniques for selecting the most suitable enhanced oil recovery method; challenges and opportunities. *Journal of Petroleum Science and Engineering*, Elsevier, v. 205, p. 108761, 10 2021. ISSN 0920-4105.

CLAPHAM, C.; NICHOLSON, J. J. R. The concise oxford dictionary of mathematics. Oxford University Press, p. 510, 2009.

COHEN, C. R. Role of fault rejuvenation in hydrocarbon accumulation and structural evolution of reconcavo basin, northeastern brazil. *AAPG Bulletin (American Association of Petroleum Geologists); (United States)*, v. 69, p. 65–76, 1985.

CORPORATION, I. B. M. *Statistics - parametric and nonparametric - IBM Documentation*. 2014. Disponível em: <https://www.ibm.com/docs/en/db2woc?topic=procedures-statistics-parametric-nonparametric>.

COUDERT, L.; FRAPPA, M.; ARIAS, R. A statistical method for litho-facies identification. *Journal of Applied Geophysics*, v. 32, n. 2, p. 257–267, 1994. ISSN 0926-9851. Disponível em: <https://www.sciencedirect.com/science/article/pii/0926985194900264>.

COURA, A. P. P. *ANÁLISE DE FÁCIES DA FORMAÇÃO MARACANGALHA (CRETÁCEO INFERIOR) NO CAMPO DE GÁS DE JACUÍPE, BACIA DO RECÔNCAVO (BA)*. Tese (Dissertação) — Universidade Federal do Rio de Janeiro (UFRJ), Rio de Janeiro, dec 2006.

CRACKNELL, M. J.; READING, A. M. Geological mapping using remote sensing data: A comparison of five machine learning algorithms, their response to variations in the spatial distribution of training data and the use of explicit spatial information. *Computers & Geosciences*, v. 63, p. 22–33, 2014. ISSN 0098-3004. Disponível em: <https://www.sciencedirect.com/science/article/pii/S0098300413002720>.

Da Silva, C. F. et al. Inversion of depocenters during the tectono-depositional evolution of the Caruaçu Member, Maracangalha Formation, in the Massapê Field, Recôncavo Basin. In: . Rio de Janeiro: SBGF (Sociedade Brasileira de Geofísica), 2019.

Da Silva, D. E. R. *CARACTERIZAÇÃO SEDIMENTOLÓGICA E ESTRATIGRÁFICA DE TESTEMUNHOS DA FORMAÇÃO ITAPARICA, CAMPO FAZENDA ALVORADA, BACIA DO RECÔNCAVO, BAHIA, BRASIL*. Tese (Monografia) — (UFBA) Universidade Federal da Bahia, Salvador, 2013.

DELFINER, P.; PEYRET, O.; SERRA, O. Automatic determination of lithology from well logs. *SPE Formation Evaluation*, v. 2, n. 03, p. 303–310, 09 1987. ISSN 0885-923X. Disponível em: <https://doi.org/10.2118/13290-PA>.

DELL'AVERSANA, P.; CIURLO, B.; COLOMBO, S. Integrated geophysics and machine learning for risk mitigation in exploration geosciences. In: *80th EAGE Conference and Exhibition 2018: Opportunities Presented by the Energy Transition*. European Association of Geoscientists and Engineers, EAGE, 2018. v. 2018, n. 1, p. 1–5. ISBN 9789462822542. Disponível em: <https://www.earthdoc.org/content/papers/10.3997/2214-4609.201801619>.

DIETTERICH, T. G. An experimental comparison of three methods for constructing ensembles of decision trees: Bagging, boosting, and randomization. *Machine Learning 2000 40:2*, Springer, v. 40, p. 139–157, 8 2000. ISSN 1573-0565. Disponível em: <https://link.springer.com/article/10.1023/A:1007607513941>.

DOMINGOS, P.; PAZZANI, M. On the optimality of the simple bayesian classifier under zero-one loss. *Machine learning*, Springer, v. 29, n. 2, p. 103–130, 1997.

Dos Santos, G. F. R. *AVALIAÇÃO PETROFÍSICA EM SISTEMAS TURBIDÍTICOS DO POÇO 7-MP-22-BA, CAMPO DE MASSAPÊ, BACIA DO RECÔNCAVO, BAHIA*. Tese (Monografia) — Universidade Federal Fluminense (UFF), Niterói, 2019.

DOWDY, S. M.; WEARDEN, S.; CHILKO, D. M. Statistics for research. Wiley-Interscience, p. 627, 2004.

DVORKIN, J.; GUTIERREZ, M. A.; GRANA, D. *Seismic reflections of rock properties*. [S.l.]: Cambridge University Press, 2014.

ELISH, M. O.; HELMY, T.; HUSSAIN, M. I. Empirical study of homogeneous and heterogeneous ensemble models for software development effort estimation. *Mathematical Problems in Engineering*, v. 2013, 2013. ISSN 1024123X.

ELKATTAN, M.; ALFY, I. A.; ELAWADI, E. Intelligent integration of neutron, density and gamma ray data for subsurface characterization. *Sensing and Imaging*, Springer, v. 21, n. 1, p. 1–14, 2020.

ELLIS, D. V.; SINGER, J. M. *Well logging for earth scientists*. [S.l.]: Springer, 2007. 692 p. ISBN 1402037384.

FENG, R.; GRANA, D.; BALLING, N. Imputation of missing well log data by random forest and its uncertainty analysis. *Computers & Geosciences*, v. 152, p. 104763, 03 2021.

FRANK, E. et al. Naive bayes for regression. *Machine Learning*, Springer, v. 41, n. 1, p. 5–25, 2000.

FREIRE, A. F. M. et al. High resolution stratigraphy using well logs to identify turbidite stages in the massapê oil field, recôncavo basin, brazil. In: *16th International Congress of the Brazilian Geophysical Society*. [S.l.: s.n.], 2019.

FREIRE, A. F. M. et al. Recognition of turbidite stages in the massapê oil field, recêncavo basin - brazil, using well logs. *Journal of Petroleum Science and Engineering*, Elsevier B.V., v. 192, p. 107279, 9 2020.

GÁMEZ, J.; RUMÍ, R.; SALMERÓN, A. Unsupervised naive bayes for data clustering with mixtures of truncated. In: . [S.l.: s.n.], 2006. p. 123–130.

GEORGIEV, G. *Has Interest in Data Science Peaked Already? | by Georgi Georgiev | Towards Data Science*. 2021. Disponível em: <https://towardsdatascience.com/has-interest-in-data-science-peaked-already-437648d7f408>.

GONÇALVES, E. et al. Prediction of carbonate rock type from nmr responses using data mining techniques. *Journal of Applied Geophysics*, v. 140, p. 93–101, 05 2017.

GORDON, A.; DESTRO, N.; HEILBRON, M. *The Recôncavo-Tucano-Jatobá Rift and Associated Atlantic Continental Margin Basins*. [S.l.: s.n.], 2017. 171–185 p.

GRUS, J. *Data Science from Scratch: First Principles with Python*. 1. ed. [S.l.]: O'Reilly Media, 2015. ISBN 149190142X,9781491901427.

HALL, M.; HALL, B. Distributed collaborative prediction: Results of the machine learning contest. *The Leading Edge*, Society of Exploration Geophysicists, v. 36, n. 3, p. 267–269, mar 2017. ISSN 1070-485X. Disponível em: <https://library.seg.org/doi/10.1190/tle36030267.1>.

HAND, D. J.; YU, K. Idiot's Bayes: Not So Stupid after All? *International Statistical Review / Revue Internationale de Statistique*, International Statistical Institute (ISI), v. 69, n. 3, p. 385, dec 2001.

HANDHAL, A. M. et al. Spatial assessment of gross vertical reservoir heterogeneity using geostatistics and gis-based machine-learning classifiers: A case study from the zubair formation, rumaila oil field, southern iraq. *Journal of Petroleum Science and Engineering*, Elsevier, v. 208, p. 109482, 1 2022. ISSN 0920-4105.

HASSAN, S.; RAFI, M.; SHAIKH, M. S. Comparing svm and naïve bayes classifiers for text categorization with wikitology as knowledge enrichment. In: *2011 IEEE 14th International Multitopic Conference*. [s.n.], 2011. p. 31–34. Disponível em: <https://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=6151495>.

HORROCKS, T.; HOLDEN, E.-J.; WEDGE, D. Evaluation of automated lithology classification architectures using highly-sampled wireline logs for coal exploration. *Computers & Geosciences*, v. 83, p. 209–218, 2015. ISSN 0098-3004. Disponível em: <https://www.sciencedirect.com/science/article/pii/S0098300415300212>.

HOSSAIN, T. M. et al. Missing well log data handling in complex lithology prediction: An nis apriori algorithm approach. *International journal of innovative computing, information & control: IJICIC*, v. 16, p. 1077–1091, 06 2020.

HURWITZ, J.; KIRSCH, D. *O que é Machine Learning e como utilizar? - IBM Brasil*. 2018. Disponível em: <https://www.ibm.com/br-pt/analytics/machine-learning>.

ISAAKS, E. H.; SRIVASTAVA", R. M. *Applied Geostatistics*. 1. ed. New York: Oxford University hess, 1989.

JIA, R. et al. A stacking methodology of machine learning for 3d geological modeling with geological-geophysical datasets, laochang sn camp, gejiu (china). *Computers and Geosciences*, Pergamon, p. 104754, mar 2021. ISSN 00983004. Disponível em: <https://linkinghub.elsevier.com/retrieve/pii/S0098300421000625>.

JOYCE, J. M. *Bayes' Theorem (Stanford Encyclopedia of Philosophy)*. 2003. Disponível em: <http://seop.illc.uva.nl/entries/bayes-theorem/>.

KEESMAN, K. *System Identification: An Introduction*. Springer London, 2011. (Advanced Textbooks in Control and Signal Processing). ISBN 9780857295224. Disponível em: <https://www.springer.com/gp/book/9780857295217>.

KOUTROUMBAS, K.; THEODORIDIS, S. *Pattern Recognition*. [S.l.]: Elsevier Science, 2008. ISBN 9780080949123.

KRONMAL, R.; TARTER, M. The estimation of probability densities and cumulatives by fourier series methods. *Journal of the American Statistical Association*, [American Statistical Association, Taylor & Francis, Ltd.], v. 63, n. 323, p. 925–952, 1968. ISSN 01621459. Disponível em: <http://www.jstor.org/stable/2283885>.

KUANG, W.; YUAN, C.; ZHANG, J. Real-time determination of earthquake focal mechanism via deep learning. *Nature Communications*, Nature Publishing Group, v. 12, n. 1, p. 1432, dec 2021. ISSN 2041-1723. Disponível em: <http://www.nature.com/articles/s41467-021-21670-x>.

KUMAR, T.; SEELAM, N. K.; RAO, G. S. Lithology prediction from well log data using machine learning techniques: A case study from talcher coalfield, eastern india. *Journal of Applied Geophysics*, v. 199, p. 104605, 2022. ISSN 0926-9851. Disponível em: <https://www.sciencedirect.com/science/article/pii/S0926985122000763>.

LAN, X. et al. Log facies identification in carbonate reservoirs using multiclass semi-supervised learning strategy. *Fuel*, v. 302, p. 121145, 2021. ISSN 0016-2361. Disponível em: <https://www.sciencedirect.com/science/article/pii/S0016236121010243>.

LANGLEY, P.; IBA, W. *An Analysis of Bayesian Classifiers*. San Jose, 1992. 6 p.

L.F., M. Deformation mechanisms in porous sandstones: Implications for development of fault seal and migration paths in the recôncavo basin, brazil. *AAPG Memoir*, American Association of Petroleum Geologists (AAPG), v. 73, p. 195 – 212, 2000. Disponível em: <https://www.scopus.com/inward/record.uri?eid=2-s2.0-33750823066&partnerID=40&md5=57adfd190e4a059ca6b5e518036a8ba5>.

LI, X.; CHAN, C. W.; NGUYEN, H. H. Application of the Neural Decision Tree approach for prediction of petroleum production. *Journal of Petroleum Science and Engineering*, Elsevier, v. 104, p. 11–16, apr 2013. ISSN 09204105.

LI, Y.; ANDERSON-SPRECHER, R. Facies identification from well logs: A comparison of discriminant analysis and naïve Bayes classifier. *Journal of Petroleum Science and Engineering*, 2006.

LINDBERG, D.; RIMSTAD, E.; OMRE, H. Inversion of well logs into facies accounting for spatial dependencies and convolution effects. *Journal of Petroleum Science and Engineering*, Elsevier, v. 134, p. 237–246, oct 2015.

LOPES, R. L.; JORGE, A. M. Assessment of predictive learning methods for the completion of gaps in well log data. *Journal of Petroleum Science and Engineering*, Elsevier, v. 162, p. 873–886, 2018.

MAGNAVITA, L. P.; SILVA, H. T. F. da. Rift border system: The interplay between tectonics and sedimentation in the reconcavo basin, northeastern brazil. *AAPG Bulletin*, v. 79, 1995.

MARR, B. *4 Mind-Blowing Ways Facebook Uses Artificial Intelligence*. 2016. Disponível em: <https://www.forbes.com/sites/bernardmarr/2016/12/29/4-amazing-ways-facebook-uses-deep-learning-to-learn-everything-about-you/?sh=3bc906a5ccbf>.

MASOUDI, P. et al. Application of bayesian in determining productive zones by well log data in oil wells. *Journal of Petroleum Science and Engineering*, v. 94-95, p. 47–54, 2012. ISSN 0920-4105. Disponível em: <https://www.sciencedirect.com/science/article/pii/S0920410512001702>.

MCKINNEY, W. *Python for Data Analysis*. [S.l.]: O'Reilly Media, 2012. ISBN 978-1449319793.

MEDVEDEV, I.; WU, H.; GORDON, T. *Powered by AI: Instagram's Explore recommender system*. 2020. Disponível em: <https://ai.facebook.com/blog/powered-by-ai-instagrams-explore-recommender-system/>.

MELLO, V. L. de; LUPINACCI, W. M. Mineralogy based classification of carbonate rocks using elastic parameters: A case study from buzios field. *Journal of Petroleum Science and Engineering*, v. 209, p. 109962, 2022. ISSN 0920-4105. Disponível em: <https://www.sciencedirect.com/science/article/pii/S092041052101576X>.

MOLA, F. Classification and regression trees software and new developments. In: RIZZI, A.; VICHI, M.; BOCK, H.-H. (Ed.). *Advances in Data Science and Classification*. Berlin, Heidelberg: Springer Berlin Heidelberg, 1998. p. 311–318. ISBN 978-3-642-72253-0.

MORADI, M.; TOKHMECHI, B.; PEDRAM, M. Inversion of well logs into rock types, lithofacies and environmental facies, using pattern recognition, a case study of carbonate Sarvak Formation. *Carbonates and Evaporites*, Springer-Verlag, v. 34, n. 2, p. 335–347, jun. 2017. Disponível em: <https://link.springer.com/article/10.1007/s13146-017-0388-8#citeas>.

MURAKAMI, Y.; MIZUGUCHI, K. Applying the Naïve Bayes classifier with kernel density estimation to the prediction of protein-protein interaction sites. *Bioinformatics*, v. 26, n. 15, p. 1841–1848, 2010. ISSN 13674803.

MURPHY, K. *Machine Learning: A Probabilistic Perspective*. [S.l.]: MIT Press, 2012. (Adaptive Computation and Machine Learning series). ISBN 9780262018029.

MURTY, M. N. M. N.; DEVI, V. S. *Pattern recognition : an algorithmic approach*. [S.l.]: Springer, 2011. 263 p. ISBN 0857294946.

MUTTI, E.; NORMARK, W. R. An integrated approach to the study of turbidite systems. Springer, New York, NY, p. 75–106, 1991. Disponível em: <https://link.springer.com/chapter/10.1007/978-1-4684-8276-8_4>.

NELIZE, L. S. *ESTUDO DOS SENTIDOS DE FLUXOS GRAVITACIONAIS DA FORMAÇÃO MARACANGALHA (EOCRETÁCEO). BOM DESPACHO, NNE DA ILHA DE ITAPARICA, BAHIA, BRASIL*. Tese (Monografia) — UNIVERSIDADE FEDERAL DA BAHIA (UFBA), Salvador, dec 2011.

NETTO, A.; OLIVEIRA, J. O preenchimento do rift - valley na bacia do reconcavo. *Revista brasileira de geociencias*, v. 15, p. 97–102, 1985.

NEWMAN, D. et al. *UCI Repository of machine learning databases*. 1998. Disponível em: <http://www.ics.uci.edu/~mlearn/MLRepository.html>.

NUALART, D. Skorokhod topology. *Encyclopaedia of Mathematics*, p. 18–19, 2011.

OLÍVIO, B. d. S. et al. *Bacia do Recôncavo*. Rio de Janeiro: [s.n.], 2007. 423 – 432 p.

ONO, H. et al. *AI4Mars — Zooniverse*. 2020. Disponível em: <https://www.zooniverse.org/projects/hiro-ono/ai4mars>.

ORS, A. O. *The Role of Machine Learning in Autonomous Vehicles | Electronic Design*. 2020. Disponível em: <https://www.electronicdesign.com/markets/automotive/article/21147200/nxp-semiconductors-the-role-of-machine-learning-in-autonomous-vehicles>.

PEDREGOSA, F. et al. *1.10. Decision Trees — scikit-learn 0.24.2 documentation*. 2011. Disponível em: <https://scikit-learn.org/stable/modules/tree.html>.

PEDREGOSA, F. et al. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, v. 12, p. 2825–2830, 2011.

PETRÓLEO BRASILEIRO S. A. *5 Coisas que você precisa saber sobre a renovação da Bacia de Campos*. 2020. Disponível em: <https://petrobras.com.br/fatos-e-dados/5-coisas-que-voce-precisa-saber-sobre-a-renovacao-da-bacia-de-campos.htm>.

PRADHAN, B. A comparative study on the predictive ability of the decision tree, support vector machine and neuro-fuzzy models in landslide susceptibility mapping using GIS. *Computers and Geosciences*, Pergamon, v. 51, p. 350–365, feb 2013. ISSN 00983004.

PRATES, I.; FERNANDEZ, R. *Sumário Geológico e Setores em Oferta*. [S.l.], 2015. Disponível em: <http://rodadas.anp.gov.br/arquivos/Round_13/areas_oferecidas_r13/Sumarios_Geologicos/Sumario_Geologico_Bacia_Reconcavo_R13.pdf>.

PUGA, J. L.; KRZYWINSKI, M.; ALTMAN, N. Points of significance: Bayes' theorem. *Nature Methods*, Nature Publishing Group, v. 12, p. 277–278, 3 2015. ISSN 15487105.

QIN, R. et al. Bayesian inversion of well logs for petrophysical properties estimation. In: . [S.l.: s.n.], 2017. p. 1067–1070.

RAJKOMAR, A.; DEAN, J.; KOHANE, I. Machine Learning in Medicine. *New England Journal of Medicine*, Massachusetts Medical Society, v. 380, n. 14, p. 1347–1358, apr 2019. ISSN 0028-4793. Disponível em: <http://www.nejm.org/doi/10.1056/NEJMra1814259>.

RAN, A. R. et al. *Deep learning in glaucoma with optical coherence tomography: a review*. Springer Nature, 2021. 188–201 p. Disponível em: <https://www.nature.com/articles/s41433-020-01191-5>.

RISH, I. An empirical study of the naive Bayes classifier. In: *IJCAI 2001 workshop on empirical methods in artificial intelligence*. [S.l.: s.n.], 2001. v. 3, n. 22, p. 41–46.

ROGERS, S. J. et al. Determination of Lithology from Well Logs Using a Neural Network. *AAPG Bulletin*, v. 76, n. 5, p. 731–739, 05 1992. ISSN 0149-1423. Disponível em: <https://doi.org/10.1306/BDFF88BC-1718-11D7-8645000102C1865D>.

ROSENBLATT, M. Remarks on some nonparametric estimates of a density function. *https://doi.org/10.1214/aoms/1177728190*, Institute of Mathematical Statistics, v. 27, p. 832–837, 9 1956. ISSN 0003-4851. Disponível em: <https://projecteuclid.org/journals/annals-of-mathematical-statistics/volume-27/issue-3/Remarks-on-Some-Nonparametric-Estimates-of-a-Density-Function/10.1214/aoms/1177728190.fullhttps://projecteuclid.org/journals/annals-of-mathematical-statistics/volume-27/issue-3/Remarks-on-Some-Nonparametric-Estimates-of-a-Density-Function/10.1214/aoms/1177728190.short>.

ROUSSEEUW, P. J. Silhouettes: A graphical aid to the interpretation and validation of cluster analysis. *Journal of Computational and Applied Mathematics*, v. 20, p. 53–65, 1987. ISSN 0377-0427. Disponível em: <https://www.sciencedirect.com/science/article/pii/0377042787901257>.

S, G. G.; L, B. Fácies sedimentares gravitacionais e deformacionais da formação maracangalha em afloramento e sua importância na exploração da bacia do recôncavo. In: *2 CONGRESSO BRASILEIRO DE P&D EM PETRÓLEO & GÁS*. Rio de Janeiro: ABPG e UFRJ, 2003. Disponível em: <http://www.portalabpg.org.br/site_portugues/anais/anais2/>.

SALES, T. S.; BIJANI, R.; FREIRE, A. F. M. 2D gravity modelling integrated with multi gephysical data to investigate the basement of the southern compartment of Reconcavo basin. In: D'AVILA, R.; CARVALHO, I. d. S. (Ed.). *49º Congresso Brasileiro*

de Geologia. Rio de Janeiro: SBG (Sociedade Brasileira de Geologia), 2019. Disponível em: <https://www.49cbg.com.br/lista-aprovados.pdf>.

SAMMUT, C.; WEBB, G. *Encyclopedia of Machine Learning*. Springer US, 2011. (Encyclopedia of Machine Learning). ISBN 9780387307688. Disponível em: <https://link.springer.com/referencework/10.1007/978-0-387-30164-8>.

SAMUEL, A. L. Some Studies in Machine Learning Using the Game of Checkers. *IBM Journal of Research and Development*, v. 3, n. 3, p. 210–229, jul 1959. ISSN 0018-8646. Disponível em: <http://ieeexplore.ieee.org/document/5392560/>.

SANTOS, N. L.; CORREA-GOMES, L. C. Study on the gravitational flow directions of the maracangalha formation (early cretaceous), bom despacho, nne area of itaparica island, bahia, brazil. *Brazilian Journal of Geology*, Sociedade Brasileira de Geologia, v. 48, p. 503–518, 7 2018. ISSN 2317-4889. Disponível em: <http://www.scielo.br/j/bjgeo/a/rrZ8yx7ZrzK8pkMn37NKk3D/?lang=en>.

SCOTT, D. W. Multivariate density estimation : Theory, practice, and visualization. Wiley, p. 381, 2015.

SIDDIG, O. et al. Real-time prediction of Poisson's ratio from drilling parameters using machine learning tools. *Scientific Reports 2021 11:1*, Nature Publishing Group, v. 11, n. 1, p. 1–13, jun 2021. ISSN 2045-2322. Disponível em: <https://www.nature.com/articles/s41598-021-92082-6>.

SILVA, C. F. da. *DISTRIBUIÇÃO ESPACIAL DE SISTEMAS E ESTÁGIOS TURBIDÍTICOS DO MEMBRO CARUAÇU DA FORMAÇÃO MARACANGALHA, NO CAMPO DE MASSAPÊ, BACIA DO RECÔNCAVO.* Tese (Dissertação) — Universidade Federal Fluminense, Niteroi, 2020.

SILVA, T. et al. ESTIMATIVA E AVALIAÇÃO DAS PROPRIEDADES PETROFÍSICAS DOS RESERVATÓRIOS DAS FORMAÇÕES ÁGUA GRANDE E SERGI DO CAMPO DE SOCORRO, BACIA DO RECÔNCAVO. In: . Rio de Janeiro: SBG (Sociedade Brasileira de Geologia), 2018.

SILVERMAN, B. W. Density estimation for statistics and data analysis. Chapman and Hall, p. 175, 1986.

SILVERMAN, B. W.; JONES, M. C. E. fix and j.l. hodges (1951): An important contribution to nonparametric discriminant analysis and density estimation: Commentary on fix and hodges (1951). *International Statistical Review / Revue Internationale de Statistique*, JSTOR, v. 57, p. 233, 12 1989. ISSN 03067734.

SINGH, U. K. Fuzzy inference system for identification of geological stratigraphy off prydz bay, east antarctica. *Journal of Applied Geophysics*, v. 75, n. 4, p. 687–698, 2011. ISSN 0926-9851. Disponível em: <https://www.sciencedirect.com/science/article/pii/S0926985111001716>.

SMOLA, A. J.; SCHÖLKOPF, B. *A tutorial on support vector regression*. 2004.

SPRENT, P. Applied nonparametric statistical methods. *Applied Nonparametric Statistical Methods*, Springer Netherlands, 1988. Disponível em: <https://link.springer.com/book/10.1007/978-94-009-1223-6>.

STRUYK, C.; KARST, J. *LAS Version 2.0: A Digital Standard for Logs*. 2014. Disponível em: <https://www.cwls.org/wp-content/uploads/2014/09/LAS_20_Update_Jan2014.pdf>.

TANG, H.; WHITE, C. D. Multivariate statistical log log-facies classification on a shallow marine reservoir. *Journal of Petroleum Science and Engineering*, Elsevier, v. 61, n. 2-4, p. 88–93, aug 2008.

TARTER, M. E.; KRONMAL, R. A. An introduction to the implementation and theory of nonparametric density estimation. *American Statistician*, v. 30, p. 105–112, 1976. ISSN 15372731.

TEWARI, S.; DWIVEDI, U. D. A comparative study of heterogeneous ensemble methods for the identification of geological lithofacies. *Journal of Petroleum Exploration and Production Technology*, Springer, v. 10, p. 1849–1868, 6 2020. ISSN 21900566. Disponível em: <chrome-extension://dagcmkpagjlhakfdhnbomgmjdpkdklff/enhanced-reader.html?pdf=https%3A%2F%2Fbrxt.mendeley.com%2Fdocument%2Fcontent%2Faf5f9731-badc-3868-87e3-be1fa2b9f5d9&doi=10.1007/S13202-020-00839-Y>.

TURING, A. *Computing Machinery and Intelligence*. Blackwell for the Mind Association, 1950. (Mind: a quarterly review). Disponível em: <https://academic.oup.com/mind/article/LIX/236/433/986238>.

UPTON, G.; COOK, I. A dictionary of statistics. *A Dictionary of Statistics*, Oxford University Press, 9 2014.

VIRTANEN, P. et al. SciPy 1.0: Fundamental Algorithms for Scientific Computing in Python. *Nature Methods*, v. 17, p. 261–272, 2020.

VRUGT, J. A.; A., J. The Scientific Method, Diagnostic Bayes, and How to Detect Epistemic Errors. *AGUFM*, v. 2015, p. H31M–03, 2015. Disponível em: <https://ui.adsabs.harvard.edu/abs/2015AGUFM.H31M..03V/abstract>.

WATSON, G. S. Density estimation by orthogonal series. *The Annals of Mathematical Statistics*, Institute of Mathematical Statistics, v. 40, n. 4, p. 1496–1498, 1969. ISSN 00034851. Disponível em: <http://www.jstor.org/stable/2239615>.

WRONA, T. et al. Seismic facies analysis using machine learning. *GEOPHYSICS*, Society of Exploration Geophysicists, v. 83, n. 5, p. O83–O95, sep 2018. ISSN 0016-8033. Disponível em: <https://library.seg.org/doi/10.1190/geo2017-0595.1>.

XIANG, Z. L.; YU, X. R.; KANG, D. K. Experimental analysis of naïve bayes classifier based on an attribute weighting framework with smooth kernel density estimations. *Applied Intelligence*, Springer New York LLC, v. 44, p. 611–620, 4 2016. ISSN 15737497. Disponível em: <https://link.springer.com/article/10.1007/s10489-015-0719-1>.

XIE, Y. et al. Evaluation of machine learning methods for formation lithology identification: A comparison of tuning processes and model performances. *Journal of Petroleum Science and Engineering*, v. 160, p. 182–193, 2018. ISSN 0920-4105. Disponível em: <https://www.sciencedirect.com/science/article/pii/S0920410517308094>.

XU, T. et al. Evaluation of active learning algorithms for formation lithology identification. *Journal of Petroleum Science and Engineering*, v. 206, p. 108999, 2021. ISSN 0920-4105. Disponível em: <https://www.sciencedirect.com/science/article/pii/S0920410521006586>.

XU, Z. et al. Integrated lithology identification based on images and elemental data from rocks. *Journal of Petroleum Science and Engineering*, v. 205, p. 108853, 2021. ISSN 0920-4105. Disponível em: <https://www.sciencedirect.com/science/article/pii/S0920410521005143>.

YADAV, K.; THAREJA, R. Comparing the performance of naive bayes and decision tree classification using r. *International Journal of Intelligent Systems and Applications*, v. 11, p. 11–19, 2019. Disponível em: <https://www.mecs-press.org/ijisa/ijisa-v11-n12/IJISA-V11-N12-2.pdf>.

YANG, B. et al. Fuzzy constrained inversion of magnetotelluric data using guided fuzzy c-means clustering. In: *SEG International Exposition and Annual Meeting 2019*. Society of Exploration Geophysicists, 2020. p. 1184–1188. Disponível em: <https://library.seg.org/doi/abs/10.1190/segam2019-3215317.1>.

ZHANG, H. The optimality of naive bayes. In: . [S.l.: s.n.], 2004. v. 2.

ZHANG, H. Exploring conditions for the optimality of naïve bayes. *International Journal of Pattern Recognition and Artificial Intelligence*, v. 19, n. 02, p. 183–198, 2005.