Predição da Litologia Utilizando Perfis de Perfuração

Eric Monteiro e Lobo Luz
Departamento de Engenharia Elétrica –
Pontifícia Universidade Católica (PUC-Rio)
Rio de Janeiro – RJ – Brasil
ericluz@tecgraf.puc-rio.com

Karla Figueiredo
Departamento de Ciências da Computação
Universidade do Estado do Rio de Janeiro (UERJ)
Rio de Janeiro – RJ – Brazil
karlafigueiredo@ime.uerj.br

Abstract- A indústria do petróleo está entre uma das atividades mais importantes e caras no mundo, atraindo muita atenção e esforço. Os custos envolvidos na exploração estimulam estudos que visam o aumento do conhecimento que promova a redução das despesas com as atividades do processo de prospecção, exploração e refino. Neste contexto, a utilização de algoritmos capazes de predizer com maior acurácia a litologia, que está estreitamente relacionada à caracterização de reservatórios, tornam-se conhecimentos valiosos. Este trabalho se insere nesse contexto e tem por objetivo a inferência da litologia na área da bacia de petróleo a partir de modelos de Machine Learning, tais como: k-Nearest Neighbors, Random Forest e Redes Neurais, utilizando um número menor de perfis de perfuração. Os resultados mostraram se promissores, podendo ser utilizados no processamento de log de poços para discriminação litológica.

Keywords—Machine Learning, Petróleo, Litologia, Redes Neurais, Random Forest, k-NN

I. INTRODUCTION

Avaliação de formações geológica visa definir em termos qualitativos e quantitativos o potencial de um campo petrolífero. O conhecimento destas características é fundamental para determinação da viabilidade econômica do poço perfurado [1].

O perfil de um poço é um dado, em relação à profundidade, de uma ou mais características ou propriedades das rochas perfuradas (resistividade elétrica, potencial eletroquímico natural, tempo de trânsito de ondas mecânicas, radioatividade natural induzida, etc.). Essas propriedades são obtidas através do deslocamento contínuo de um sensor de perfilagem (sonda) dentro do poço. Os perfis mais comuns são: log de *gamma ray* (GR) registra intensidade de fonte radioativa (minerais argilosos como componente majoritário) apresentada na composição mineralógica da seção de rocha; log de densidade (ROB) é usado para medir a densidade em massa; log de nêutrons (DPHI) mede o número de poros na seção perfurada; log de resistividade (RL) é usado para distinguir a natureza do fluido nas formações geológicas; log de acústico é usado para medir a rigidez da rocha medindo a velocidade das ondas sônicas na seção perfurada. Segundo Fanchi, J. R. et al. [1], os perfis mais utilizados para determinação do litologia são o *gamma ray* (GR), registro espontâneo (SP), resistividade (RL) e acústico (DT). Estes perfis apresentam diferentes custos de aquisição, o que torna interessante inferir a litologia com menor número de perfis.

A identificação manual de litologias é um processo demanda uma quantidade considerável de tempo de um especialista experiente. Uma ferramenta precisa que automatize e acelere este processo permitirá a que este especialista dedique seu tempo para tarefas com maior grão de importância.

Wang et al. (2018) [2] utilizaram o algoritmo de agrupamento K-NN para agrupar perfis de perfuração comparando com as nove litologias presente no campo de Gaoqing. Neste trabalho foram obtidas taxas de precisão de 60 a 100 % dependendo da litologia analisada.

No trabalho de Imamverdiyev et al. (2019) [3] foi utilizada uma rede neural artificial profunda convolucional unidimensional (1D-CNN) para predizer a litologia. As variáveis de entrada foram o efeito fotoelétrico (PE), GR, registro de resistividade (RL), diferença de porosidade de densidade de nêutrons (DPHI), porosidade de densidade média de nêutrons (PHIA). Com esta estratégia obteve-se sucesso em classificar a litologia com precisão de aproximadamente 95% no seu melhor modelo.

Este presente trabalho tem como objetivo analisar a aptidão dos algoritmos de *Machine Learning*: *k-Nearest-Neighbor* (k-NN) [7] e *Random Forest* (RF) [8] e modelos baseados em Redes Neurais (RN) [4] para predição de litologia a partir de perfis de perfuração. Neste trabalho, após investigação dos perfis disponíveis, optou-se pela utilização dos perfis *gamma ray*, densidade e acústico visando obter a melhor a precisão com menor custo de aquisição.

Os dados utilizados neste trabalho encontram-se disponíveis publicamente em: https://www.nlog.nl/kaart-boringen. Este site fornece informações sobre exploração e produção de petróleo, gás e energia geotérmica do setor holandês da plataforma continental do Mar do Norte.

O restante desse trabalho está distribuído em mais cinco seções. A seção II apresenta alguns fundamentos voltados para os algoritmos de Machine Learning. A seção III apresenta a metodologia adotada nesse trabalho. A quarta seção exibe os resultados obtidos e, finalmente, a última seção encerra o trabalho apresentando as conclusões e perspectivas de novos trabalhos.

II. Modelos e Algoritmos De Machine Learning

A. Redes Neurais Artificiais

As redes neurais artificiais são modelos computacionais inspirado nos neurônios biológicos e na estrutura massivamente paralela do cérebro, com capacidade de adquirir, armazenar e utilizar conhecimento experimental.

Em geral, a resposta de um neurônio é expressa por $y = f(w_i * x_i + \theta)$, onde uma função de ativação "f" é inferida pelo resultado da soma ponderada das entradas (x_i) pelos pesos sinápticos (w_i), acrescido pelo bias (θ). Essa função de ativação acrescenta características não lineares e limita a amplitude da saída do neurônio.

O processamento neural pode ser dividido em duas fases: (a) *Learning*, processo pelo qual os parâmetros livres, os pesos sinápticos w_i e os bias, de uma rede neural são ajustados através de um processo contínuo de estimulação pelo treinamento. Nesta etapa ocorre a aquisição da informação; (b) *Recall*, processo de inferência de dados de entrada sobre a função aprendida. Nesta etapa ocorre a recuperação da informação.

Os algoritmos de treinamento supervisionado minimizam o erro, a diferença entre a saída da rede e o valor desejado, atualizando os pesos sinápticos e bias através de várias iterações (épocas ou ciclos de treinamento).

A estrutura de uma rede neural típica é composta por uma rede de neurônios artificiais, não lineares (função de ativação), dispostos em camadas e interconectados através de canais unidirecionais, análogos às sinapses de um neurônio biológico [4].

B. k-Nearest Neighbors (k-NN)

Este é um algoritmo de aprendizagem supervisionado baseado em instância. O k-NN está entre os mais simples algoritmos de *Machine Learning*. Ele utiliza a similaridade entre os k vizinhos mais próximos para classificar o novo registro. Naturalmente, o desempenho deste algoritmo é seriamente afetado pelo tamanho da vizinhança definida pelo parâmetro k [5][6].

Este algoritmo não possui um treinamento semelhante ao das Redes Neurais, mas deve se usar parte dos dados de treinamento para avaliar e indicar os melhores parâmetros: número de vizinhos e tipo de distância para cálculo da vizinhança. Para classificação, o algoritmo calcula a distância n-dimensional entre o novo registro e os dados históricos (treinamento). A classe dos k vizinhos mais próximos (por maioria) será a classe desse novo registro [7].

Um dos problemas na utilização do k-Nearest Neighbors é está relacionado à dimensionalidade do conjunto de dados. Tal problema pode ser contornado com agrupamento prévio dos dados, limitando a busca aos registros que estiverem no grupo com centroide próximo ao novo dado a ser avaliado. Além disso, como todo método baseado em distâncias, atributos irrelevantes podem corromper a classificação [7].

C. Random forest (RF)

Este é um algoritmo de Machine Learning para aprendizagem supervisionada que se caracteriza como um *ensemble,* pois cria várias árvores de decisão e as combina para obter uma predição normalmente com maior acurácia e robustez, inclusive com viés estatístico se um número grande de árvores for usado. O algoritmo considera um conjunto de árvores de decisão, em que cada árvore é criada a partir de algoritmo baseado na entropia da informação para escolha de atributos para compor a árvore, considerando inicialmente uma seleção aleatórias de atributos.

Uma vez que cada árvore usa um subconjunto dos dados para testar o desempenho do modelo treinado. O RF possui como parâmetros, o número de profundidade máxima, número total de árvores, além de permitir a possibilidade de poda.

As arquiteturas das árvores são determinadas a partir do melhor particionamento dos dados em relação aos atributos selecionados (numérico ou categórico). A seleção de atributos que irão compor a estrutura da árvore utiliza métricas como Índice *Gini* ou taxa de ganho, entre outras métricas baseadas em entropia. O processo de divisão continua recursivamente até que um critério de parada seja alcançado (profundidade ou acurácia esperada).

O método *Random Forest* é essencialmente uma coleção de árvores de decisão, cujos resultados individuais podem ser contabilizados, agregados ou ponderados e agregados, consolidando uma resposta final [8][9] dos modelos individuais, como em um comitê. Assim, este método é considerado uma coleção de modelos (*ensemble*).

III. Metodologia

Os logs de perfuração dos poços selecionados para este estudo consideraram os perfis gamma ray (GR), acústico (DT), densidade (RHOB), densidade corrigida (DROB) e neutro (NPHI). Os perfis DROB e NPHI foram excluídos previamente devido à baixa variância e grande quantidade de dados faltantes para os dois poços usados na pesquisa: 75% e 78% dos dados faltantes, respectivamente.

Assim, os perfis que restaram nessa primeira seleção de variáveis foram o gamma ray, acústico e densidade, sendo organizados em janelas de profundidade, de modo a transmitir a dinâmica da composição espacial da litologia para os modelos de *Machine Learning*. A técnica de "janelamento", usa sequências de dados históricos (profundidade) da variável sobre a qual se deseja realizar a previsão, ou seja, para inferir o dado em p+1, utilizam-se como dados de entrada o valor da variável em p mais os p-n valores anteriores, onde n representa o número de amostras da série pertencentes à janela. Assim, cada registro da base de dados foi composto por n janelas de 2 metros com um deslocamento de 0.5 metros entre elas. Para cada janela calculou-se diferente métricas sobre esses dados: média, mediana, máximo, mínimo, média harmônica, média geométrica e soma (Fig. 1). O software "*GetData Graph Digitizer*"[11] foi utilizado para digitalizar a litologia em função da profundidade dos dois poços, pois os dados não estavam disponíveis digitalmente.

As métricas e tamanho de janelas foram avaliadas numa seleção de variáveis, utilizando o tipo *wrapper* de processo de seleção de variáveis, com modelos de inferência e a litologia conhecida como saída, visando a escolha dos melhores atributos, normalizados na faixa de -1 a +1. Nesta busca utilizou-se o software "*Orange*" [10]. Para esse processo de seleção *wrapper* avaliou-se a resposta dos algoritmos de *Machine Learning k-Nearest Neighbors* (k=3, distância euclidiana e pesos baseado nesta distancia), *Random Forest* (10 árvores, sem limite de profundida e grupo mínimo de divisão de 10) e Redes Neurais Artificiais (20 neurônios, tangente hiperbólica como função de ativação) para as diferentes combinações de atributos e tamanhos de janelas. Para este teste, utilizou-se os dados de um poço (poço 01 – F2_01) contendo as litologias *claystone*, *siltstone*, *coal*, *sandstone*, *shale* e *limestone*, nas proporções 35,82%, 0,59%, 0,11%, 15,34%, 41,61% e 6,54%, respectivamente. Os perfis *gamma ray*, densidade e sônico foram analisados separadamente por cada técnica (k-NN, *Random Forest* e Redes Neurais).

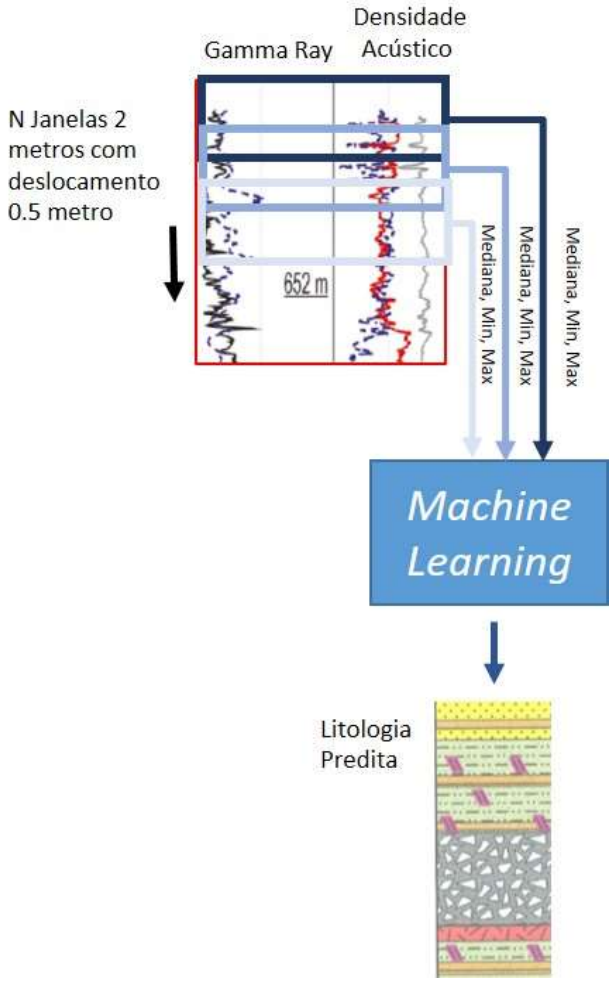

Fig. 1: Metodologia Geral para criação dos registros do banco de dados

Com as dimensões de cada janela em profundidade, as métricas calculadas sobre essa janela e perfil envolvido na avaliação, realizou-se a classificação final e o ajuste dos parâmetros dos modelos de classificação no software "Orange" versão 3.21[10]. Após o ajuste dos modelos com dados do poço 01(F2_01), estes modelos foram testados com acréscimo dos dados do poço 02 (F2_03) pertencente à mesma região. Os modelos serão avaliados com base na sua acurácia e no F1, que é a média harmônica da precisão e recall e visa trazer um número único, que indique a qualidade geral do modelo, mesmo para problemas que tenham classes desbalanceadas, conforme é o caso desse trabalho.

IV. Resultados E Discussões

Conforme mencionado, os dados foram obtidos da bacia do Mar do Norte. A base de dados possuía 5 perfis: os perfis gamma ray (GR), acústico (DT), densidade (RHOB), densidade corrigida (DROB) e neutro (NPHI). Os perfis DROB e NPHI foram excluídos. O número de registros para cada perfil foi criado com os dados coletados entre 260 metros e 3200 metros de profundidade, a partir do subsolo marinho, abaixo da lâmina d'água. Nesse caso, fez-se um pareamento eliminando os valores acima de 260 metros e abaixo de 3200 metros, do subsolo marinho, visando uniformizar os valores das variáveis disponíveis para que todas tivessem uma profundidade comum. A quantidade de classes litológicas presentes no banco de dados referentes aos poços 01 (F2_01) e 02 (F2_01) encontram-se na Fig. 2. Conforme pode se observar há enorme desbalanceamento da base de dados utilizada.

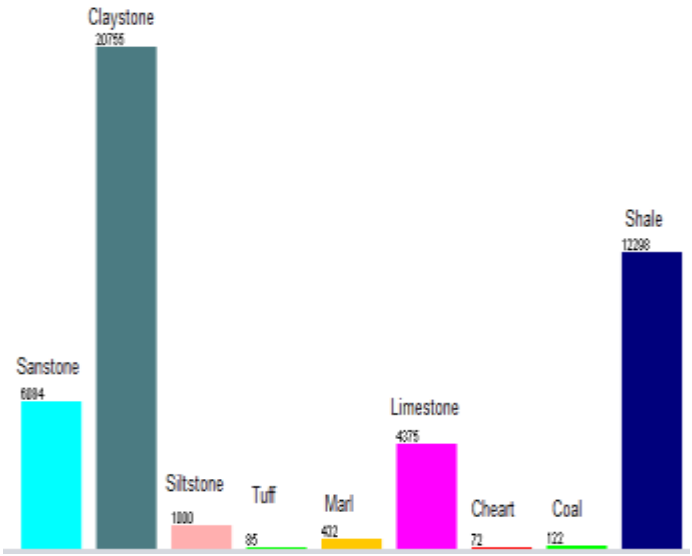

Fig. 2. Histograma do banco de dados utilizado neste trabalho

Para esses poços (Fig. 2) as litologias são: *Cheat, Claystone, Coal, Limestone, Marl, Sandstone, Siltstone, Tuff e Shale.*

Utilizando a metodologia de seleção de variável do tipo *wrapper* determinou-se as melhores métricas para os perfis de *gamma ray*, acústico e densidade, respectivamente nas Tabelas I, II e III. As métricas máximo, mínimo e mediana foram capazes de transmitir com melhor acurácia a dinâmica composição espacial dos perfis de perfuração.

Nas Tabelas I, II e III a primeira coluna indica o número da janela, a segunda coluna indica a terceira métrica utilizada, além do mínimo e máximo no janelamento. Nesta coluna: ME representa a média aritmética, MG representa a média geométrica, MH representa a média harmônica, MD representa mediana e Sum representa o somatório. As colunas 3, 4 e 5 indicam a acurácia obtida pelos modelos RNA, k-NN e RF, respectivamente.

TABELA I. RESULTADOS DA SELEÇÃO DE PARÂMETROS E VARIÁVEIS RELATIVA AO ATRIBUTO GAMMA RAY

Janela	Métricas	**Acurácia**		
		RNA	k-NN	RF
4	-	0,832	0,971	0,968
4	ME	0,832	0,973	0,960
4	**MD**	**0,837**	**0,974**	**0,969**
4	MG	0,829	0,971	0,960
4	MH	0,832	0,972	0,961
4	Sum	0,832	0,973	0,961
3	MD.	0,828	0,953	0,958

TABELA II. RESULTADOS DA SELEÇAO DE PARÂMETROS E VARIÁVEIS RELATIVA AO ATRIBUTO DENSIDADE

Janela	Métricas	**Acurácia**		
		RNA	k-NN	RF
4	-	0,807	0,962	0,952
4	ME	0,826	0,941	0,945
4	**MD**	**0,825**	**0,971**	**0,959**
4	MG	0,826	0,966	0,942
4	MH	0,825	0,969	0,944
4	Sum	0,827	0,921	0,944
3	MD.	0,799	0,942	0,938
2	MD.	0,786	0,858	0,874

TABELA III. RESULTADOS DA SELEÇÃO DE PARÄMETROS E VARIÁVEIS RELATIVA AO ACÚSTICO

Janela	Métricas	**Acurácia**		
		RNA	k-NN	RF
4	-	0,788	0,971	0,971
4	ME	0,792	0,972	0,967
4	**MD**	**0,794**	**0,973**	**0,973**
4	MG	0,791	0,970	0,966
4	MH	0,788	0,972	0,967
4	Sum	0,792	0,972	0,966
3	MD.	0,782	0,956	0,936
2	MD.	0,760	0,894	0,924

A redução das quantidades de janelas resulta em perda de acurácia nos três modelos de *Machine Learning* utilizados nesta seleção. Dos limites de parâmetros testados, a quantidade ideal foi 4 janelas de 2 metros com um deslocamento sobreposto de 0.5 m para todos os perfis (Tabelas I, II e III).

A Fig. 3 indica a estrutura dos atributos de entrada por perfil para os algoritmos e modelo de Machine Learning. Ao todo foram 36 entrada de dados, ou seja, um vetor contendo 36 valores construído a partir dos valores capturados por perfil nas janelas de 2m e calculadas as métricas indicadas na figura: min: mínimo, max: máximo e med: mediana.

gamma ray

janela 1 0m a 2m			janela 2 0,5m a 2,5m			janela 3 1m a 3m			janela 4 1,5 m a 3,5 m		
min	max	med	min	max	med	min	max	med	min	max	med

densidade

janela 1 0m a 2m			janela 2 0,5m a 2,5m			janela 3 1m a 3m			janela 4 1,5 m a 3,5 m		
min	max	med	min	max	med	min	max	med	min	max	med

acústica

janela 1 0m a 2m			janela 2 0,5m a 2,5m			janela 3 1m a 3m			janela 4 1,5 m a 3,5 m		
min	max	med	min	max	med	min	max	med	min	max	med

Fig. 3. Atributos de entrada dos modelos

As tabelas IV, V e VI apresentam os resultados obtidos a partir da acurácia e F1, calculados sobre os dados de teste, os quais foram criados a partir dos perfis do poço 2 (F2_03). Esta base de dados é composta por 45223 registros criados segundo os resultados registrados nas Tabelas I, II e III: valores mínimo, máximo e mediana obtidos por meio de métricas nas 4 janelas de 2 metros com deslocamento de 0,5 metros para os atributos acústico, densidade e *gamma ray.*

No ajuste de parâmetros dos modelos de predição da litologia utilizou-se a técnica da validação cruzada com k-*folds*, para k=5. Assim, nesse procedimento a base de dados foi dividida em cinco partes e utilizou-se 4/5 partes (4-*folds* - 36178 registros) dos dados para o treinamento, deixando o conjunto complementar (1-*fold* = 1/5 dos dados = 9045 registros) para teste. Este procedimento proporciona a construção de 5 modelos ajustados alternando a combinação das partes que compõem o treinamento e teste, gerando 5 conjuntos de teste distintos. Com isso foi possível a avaliação robusta, obtida a partir da média das métricas de avaliação para os 5-folds de teste (1-*fold*).

A Tabela IV apresenta os resultados obtidos por diferentes arquiteturas de redes neurais. Nela destaca-se que foram avaliados distintos números de neurônios na primeira ou primeira e segunda camadas escondidas, além dos algoritmos de aprendizado utilizado os algoritmos Adam (considera momentos de primeira e segunda ordem do gradiente para otimização das funções objetivo com taxas independentes para atualização dos neurônios) e L-BFGS-B (da família dos métodos quasi-Newton com uso limitado de memória). A topologia de duas camadas escondidas contendo 100 e 100 neurônios, respectivamente, obteve a melhor acurácia (0,991).

A matriz de confusão na Fig. 4 mostra os resultados para o modelo RNA (100-100 neurônios nas camadas escondidas). Nela se observa uma quantidade de registros fora da diagonal principal, ou seja, dados não classificados corretamente. As litologias com menor quantidade de acertos foram *claystone*, *siltstone* e *sandstone*. A dupla com maiores erros de classificação foram a *sandstone* e *claystone*, porém a quantidade instâncias classificadas incorretamente é muito menor considerando a quantidade total de instancias.

TABELA IV. RESULTADOS RELATIVOS AO AJUSTE DE PARÂMETROS DA REDE NEURAL ARTIFICIAL

Nº de Neurônios	**Treinamento**	**Acurácia**	**F1**
1	L-BFGS-B	0,731	0,671
5	L-BFGS-B	0,824	0,813
10	L-BFGS-B	0,832	0,840
15	L-BFGS-B	0,870	0,873
20	L-BFGS-B	0,868	0,872
50	L-BFGS-B	0,930	0,894
100	L-BFGS-B	0,921	0,922
75	L-BFGS-B	0,918	0,919
75	Adam	0,927	0,927
100	Adam	0,933	0,933
100-25	Adam	0,966	0,966
100-50	Adam	0,979	0,979
100-100	**Adam**	**0,991**	**0,990**
125-100	Adam	0,990	0,990
125-125	Adam	0,990	0,990

A Tabela V apresenta os valores dos parâmetros dos modelos avaliados nos experimentos com k-NN. A coluna 1 indica o número de vizinhos próximos à amostra avaliada; a segunda coluna aponta a métrica; a coluna peso mostra a ponderação utilizada das distancias; as duas últimas colunas exibem os resultados obtidos com a base de teste para a precisão e para o F1.

Conforme pode-se perceber na Tabela V (com destaque para o modelo escolhido), para o modelo do k-NN, a distância de *manhatan* apresentou melhor acurácia independentemente do número de vizinhos. A utilização de pesos baseado em distância melhorou o desempenho do algoritmo em geral.

Analisando a matriz de confusão da Fig. 5, percebe-se que a maiorias das instâncias também se encontram-se na diagonal principal e novamente a dupla com maiores erros de classificação foram a *sandstone* e *claystone*.

A Tabela VI, com destaque para o modelo escolhido, apresenta os valores dos parâmetros dos modelos avaliados nos experimentos usados com Random Forest. A coluna 1 indica o número de árvores; a segunda e terceira colunas apontam o nível mínimo e máximo de uma árvore, respectivamente; as duas últimas colunas exibem os resultados obtidos com a base de teste para a acurácia e para o F1.

Saída do Modelo de Rede Neural

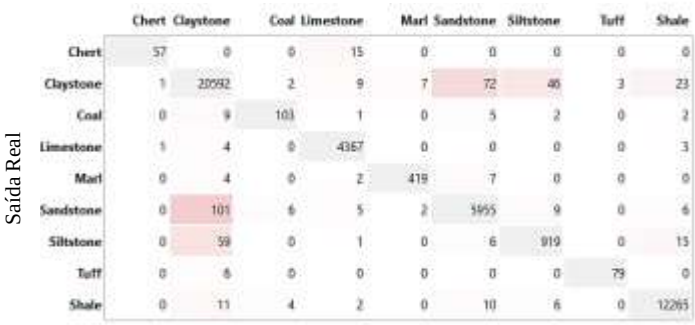

Fig. 4. Matriz de Confusão para RNA

TABELAV. RESULTADOS RELATIVOS AO AJUSTE DE PARÂMETROS DO K-NN

(k)	**Métrica**	**Peso**	**Precisão**	**F1**
3	Euclidiana	Distância	0,992	0,992
3	Manhatan	Distância	0,996	0,996
3	Chebyshev	Distância	0,961	0,961
4	Euclidiana	Distância	0,993	0,993
4	Manhatan	Distância	0,996	0,996
4	Chebyshev	Distância	0,961	0,962
5	Euclidiana	Distância	0,989	0,989
5	Manhatan	Distância	0,996	0,996
5	Chebyshev	Distância	0,947	0,947
6	Manhatan	Distância	0,995	0,995
3	Euclidiana	Uniforme	0,991	0,991
3	Manhatan	Uniforme	0,996	0,996
3	Chebyshev	Uniforme	0,956	0,956
3	Manhatan	Uniforme	0,994	0,994

Saída do Modelo k-NN

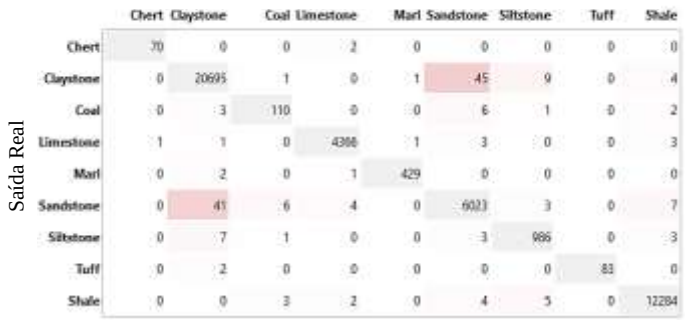

Fig. 5. Matriz de Confusão para K-NN

Para o algoritmo do *Random Forest*, a quantidade de 50 árvores com no máximo 25 níveis profundidade apresentou melhor precisão. Um dos parâmetros a ser definido é o menor subconjunto que pode ser dividido, este parâmetro impacta na complexidade da árvore e nos resultados. Sem essa limitação, permitiu ao algoritmo alcançar o F1 de 0,995 com custo de árvores mais complexas.

TABELA VI. RESULTADOS RELATIVOS AO AJUSTE DE PARÂMETROS DO RANDOM FOREST

Árvores	Gr. Mín	Nível Máx	Acurácia	F1
10	5	-	0,989	0,989
20	5	-	0,993	0,993
100	5	-	0,994	0,994
50	10	-	0,991	0,991
50	2	-	0,995	0,995
50	2	10	0,926	0,913
50	**2**	**25**	**0,995**	**0,995**
50	2	25	0,992	0,992

Através da matriz de confusão da Fig.6, percebe-se que as classes com maiores erros de classificação foram a *sandstone*, *siltstone* e *claystone*.

Saída do Modelo *Random Forest*

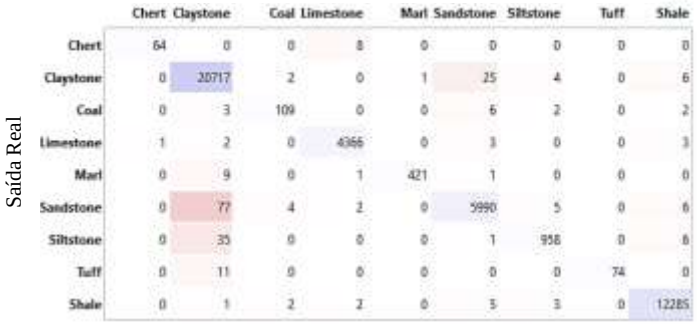

Fig. 6. Matriz de Confusão para o Random Forest

Com o índice de similaridade de Jaccard [12] (Tabela VII) comprovou-se alto grau de similaridade entre os resultados obtidos pelos modelos.

Tabela VII. ÍNDICE DE SIMILARIDADE DE JACCARD

	Rede Neural	Random Forest	kNN
Rede Neural	-	0,989	0,990
Random Forest	-	-	0,995
kNN	-	-	-

V. CONCLUSÕES

Neste trabalho avaliou-se a capacidade preditiva de diferentes modelos de *Machine Learning* para classificação de litologia utilizando poucos perfis de perfuração. A transformação dos dados em janelas móveis possibilitou transmitir a dinâmica do perfil para o modelo de *Machine Learning*, ajudando a alcançar os resultados obtidos. Os modelos de Rede Neural, *Random Forest* e k-NN apresentaram bons resultados, com F1 de 99,1%, 99,5% 99,6%, respectivamente. O índice de Jaccard indica a grande similaridade entre os resultados obtidos entre os modelos, considerando as bases teste correspondentes aos 5 *folds*. As litologias com os maiores graus de confusão foram *sandstone*, *siltstone* e *claystone*. O modelo baseado em k-NN apresentou menor grau de confusão entre todas litologias, inclusive entre *sandstone* e *claystone*.

Como perspectivas de continuidade dos trabalhos pode-se indicar diferentes tamanhos de janelamento utilizando os valores das variáveis sem qualquer processamento (métricas) e baleanceamento de dados entre as litologias. Outra investigação interessante é a avaliação da distribuição da acurácia pelos algoritmos utilizados, visando o uso de ensembles.

REFERENCES

[1] A. Frank, H-U. Krieger, F. Xu, H. Uszkoreit, B. Crysmann, B. Jörg and U. Schäfer, Well Logging, Shared Earth Modeling, 52–68, 2002.

[2] X. Wang, S. Yang, Y. Zhao, and Y. Wang, "Lithology identification using an optimized KNN clustering method based on entropy-weighed cosine distance in Mesozoic strata of Gaoqing field, Jiyang depression" Journal of Petroleum Science and Engineering, 166, 157–174. 2018.

[3] Y. Imamverdiyev and L. Sukhostat, "Lithological facies classification using deep convolutional neural network", Journal of Petroleum Science and Engineering, 174, 216–228, 2019.

[4] S. Haykin, Redes neurais: princípios e prática. trad. Paulo Martins Engel. - 2.ed. - Porto Alegre: Bookman, 2001.

[5] J. Han, J. Pei and M. Kamber, Data Mining: Concepts and Techniques, Elsevier, 2011.

[6] E. Frank and I.H. Witten, Generating Accurate Rule Sets Without Global Optimization. In: Fifteenth International Conference on Machine Learning, p. 144–151, 1998.

[7] N.S. Altman, "An Introduction to Kernel and Nearest-Neighbor Nonparametric Regression", The American Statistician, v. 46, n. 3, p. 175–185, 1992 .

[8] L. Rokach and O. Maimon, Data mining with decision trees: theory and applications, World Scientific Pub Co Inc., 2008.

[9] J.R. Quinlan, "Induction of Decision Trees", Machine Learning, v. 1, n. 1, p. 81–106, 1989.

[10] J. Demšar, A. Erjavec, T. Hočevar, M. Milutinovič, M. Možina, M. Toplak, B. Zupan, Orange: Data Mining Toolbox in Python. Journal of Machine Learning Research, 14, 2349–2353, 2013.

[11] GetData Graph Digitizer. Site: http://getdata-graph-digitizer.com/index.php. Acessado em: 12/09/19

[12] L. Hamers, Y. Hemeryck, G. Herweyers, M. Janssen, H. Keters, R. Rousseau, A. Vanhoutte Similarity measures in scientometric research: The Jaccard index versus Salton's cosine formula. Information Processing & Management, v. 25, n. 3, p. 315–318, 1989.