11º CONGRESSO BRASILEIRO DE PESQUISA E DESENVOLVIMENTO EM PETRÓLEO E GÁS

Título do Trabalho:

Classificação de litofáceis baseada em ferramentas de Aprendizagem de Máquina (AM) paramétrica e não-paramétrica

Autores:

Marcus L. A. do Amaral (FAGEOF-UFPA), Isaac N. da S. Macedo (FAGEOF-UFPA), Jose F. V. Gonçalves (FAGEOF-UFPA), Jose J. S. de Figueiredo (CPGf & FAGEOF-UFPA & INCT-GP) e Celso R. L. de Lima (CPGf-UFPA)

Instituição:

Universidade Federal do Pará
Instituto Nacional de Ciência e Tecnologia-Geofísica do Petróleo

Classificação de litofáceis baseada em ferramentas de Aprendizagem de Máquina (AM) paramétrica e não-paramétrica

Resumo

Atualmente, o uso de ferramentas de aprendizagem de máquina (AM) supervisionada e não supervisionada na classificação de fácies litológicas, tem se tornado recorrente na geofísica de poço e petrofísica. Neste trabalho, utilizamos métodos paramétricos k-Nearest Neighbors (kNN) e Logistic Regression (LR), e não paramétricos Multi Layer Perceptron (MLP), Suport Vector Machine (SVM) e Gaussian Process (GPR) na classificação de fácies a partir de dados de poços. Para testar a perfomace de cada modelo de classificação, usamos as métricas de acurácia, F1-score, precisão e recall. Também foi feita uma análise baseada na busca de hiperparâmetros para cada modelo. Baseado nos resultados em um poço cego, os métodos não paramétricos com otimização dos hiperparâmetros foram aqueles que obtiveram melhor rendimento na classificação.

Abstract

Currently, the use of supervised and unsupervised machine learning (ML) tools in lithological facies classification has become a recurrent feature in well logging and petrophysics. In this paper, we use parametric methods k-Nearest Neighbors (kNN) and Logistic Regression (LR), and nonparametric Multi Layer Perceptron (MLP), Support Vector Machine (SVM) and Gaussian Process (GPR) in facies classification from well data. To test the perfomace of each classification model, we used the metrics of accuracy, F1-score, precision and recall. A hyperparameter search-based analysis was also performed for each model. Based on the results in a blind well, the nonparametric methods with hyperparameter optimization were the ones that obtained the best classification performance.

Introdução

Procedimentos de classificação automática podem ser definidos como as ferramentas utilizadas para identificar a qual grupo um "conjunto de entrada" pertence. Do ponto de vista geofísico e geológico, classificação de fácies é importante pois, dentre vários benefícios, está a classificação automática de fácies e litofáceis de poços reais disponibilzados pela SEG (Society Exploration of Geophisics) usando métodos supervisionado e não supervisionados [2]. Aqui usamos os métodos supervisionados: Multi Layer Perceptron (MLP), Suport Vector Machine (SVM), Logistic Regression (LR), k-Nearest Neighbors (kNN) e Gaussian Process (GPR) com kernels. Os Kernel utilizados foram: o Radial Basis Fuction (RBF) e o Rational.

Neste trabalho fizemos a implementação dos métodos de classificação já mencionados com e sem a hiperparametrização. Para isso fizemos uso do GridSearch para buscar tais hiperparâmetros, e após ser encontrados foram implementado em dados cegos. Ou seja, dados que não fizeram parte de nenhuma etapa do pré-processamento e processamento. A partir de uma análise quantitativa e qualitativa os métodos não paramétricos com otimização dos hiperparâmetros apresentaram melhor rendimento na classificação do conjunto de dados referente ao campo de gás Hugoton e Panoma, no Kansas - EUA [1].

Metodologia

Para ser feita a aplicação das técnicas de classificação utilizamos a linguagem python com as bibliotecas Scikit-learn, Pandas e Numpy, essas são responsáveis pela aplicação das técnicas de AM, vizualização dos dados de Poços disponibilizados pela SEG e o tratamento de vetores e matrizes. Como ferramenta de visualização dos dados

utilizou-se o seaborn e matplotlib. Anterior a qualquer aplicação, como pode ser visto na Figura 1, é necessário fazer o carregamento dos dados que contém vários poços concatenados.

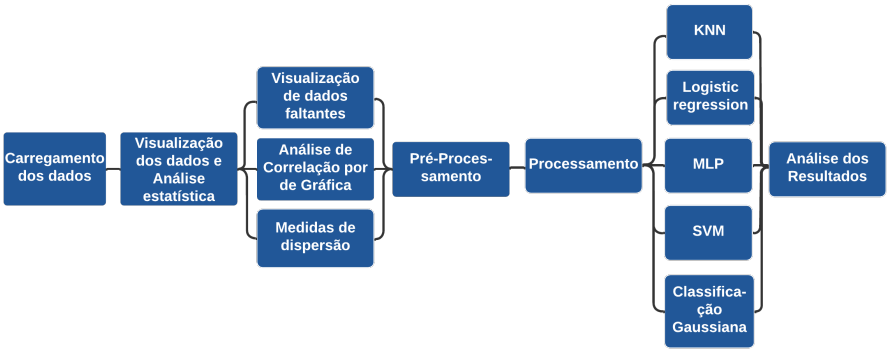

Figura 1: Fluxograma com as etapas do trabalho que começa com o carregamento dos dados seguido, pela visualização e análise, medida de dispersão, análise de correlação gráfica e visualização, pré-processamento, processamento com SVM, MLP, KNN e Logistic Regression, classificação Gaussiana e por fim com a análise dos resultados.

De acordo com os dados disponibilizados pela SEG esses são constituídos de nove poços que foram rotulados com um tipo de fácies baseado na observação do testemunho, sabendo que essas nove fáceis são descritas com seus respectivos rótulos de fáceis, legenda e fáceis adjacentes que pode ser visto na Tabela 1.

Tabela 1: Tabela que contém a descrição das rochas com seu rótulo de fácies, legenda e fácies adjacentes.

Rocha	Rótulos de Fácies	Legenda	Fácies Adjacentes
Arenito não marinho	**1**	**SS**	**2**
Siltstone grosseiro não marinho	**2**	**CSiS**	**1,3**
Siltstone fino não marinho	**3**	**FSiS**	**2**
Siltstone marinho e xisto	**4**	**SiSh**	**5**
Mudstone (calcário)	**5**	**MS**	**4,6**
Wackestone (calcário)	**6**	**WS**	**5, 7 , 9**
Dolomite	**7**	**D**	**6, 8**
Packstone-grainstone (calcário)	**8**	**PS**	**6, 7, 9**
Defletor de algas filoides (calcário)	**9**	**BS**	**7, 9**

Ao término das etapas iniciais, carregamento e análise de dados, foi feito o pré-processamento, onde dados faltantes foram cortados, análise de correlação e separação entre dados de treino e dados de teste, respectivamente 80% treino e 20% teste. Como parâmetros de entrada usou-se a formação, nome do poço, profundidade, fácies e fácies rotuladas com objetivo de prever as fácies de poços cegos, sendo que esses não fizeram parte de nenhuma das etapas. No processamento os métodos usados neste trabalho para classificação foram kNN, SVM, Regressão Logística, Classificação Gaussiana e MLP, ademais a cada método fora aplicado a hiperparametrização, ou seja, buscou-se melhores parâmetros para a melhor obtenção dos resultados e para isso utilizou-se o GridSearch.

O 'GridSearchCV' consegue automatizar etapas repetitivas do processo de tunning [3], ou seja aplicar mudanças visando otimizar o desempenho na recuperação ou atualização dos dados. Ele ajusta os hiperparâmetros de maneira sistemática com diversas combinações, após avaliá-los e armazená-los em um único objeto. O seu principal objetivo é criar combinações de hiperparâmetros, obter os melhores valores para possuir resultados de previsão excelentes. Por esse motivo o processamento é demorado e caro, com base nos hiperparâmetros envolvidos.

A Regressão logística busca prever os dados a serem classificados com o seu valor pertencente ao conjunto [0,1], tendo em vista que l é um limiar- costumeiramente igual a 0.5, sendo sua fórmula:

$$\hat{y} = \begin{cases} 1, p \geq l \\ 0, p < l. \end{cases} \quad (1)$$

Para exemplificarmos o seu funcionamento em uma classificação binária[3], onde se a determinada categoria está abaixo do limiar essa é classificada como zero e acima do limiar é classificada como 1, assim definimos a Regressão Logística como:

$$g(X) = \frac{e^{\beta_0+\beta_1 X}}{1+e^{\beta_0+\beta_1 X}}. \tag{2}$$

Com o uso do GridSearch encontrou-se para a regressão logistica o solver $newton-cg$, que é usado para problemas multiclasses e lida com perda multinomial; para a penalidade o parâmetro encontrado fora o $l2$ e parâmetro C, que tem por objetivo encontrar o inverso da força de regularização, o seu valor foi 0.1.

No entanto, k-Nearest Neighbors ou k vizinhos mais próximos, tem por objetivo determinar a classificação por meio de amostras vizinhas advindas de um conjunto de treinamento, sendo formulado da seguinte forma:

$$\hat{y}(x) = \frac{1}{k} \sum_{x_i in N_k(x)} y_i. \tag{3}$$

no qual $N_k(x)$ é a vizinhança do ponto x definida pelos k vizinhos mais próximos de x_i que são classificados como y_i. Com o GridSearch fora encontrado como melhores hiperparâmetros o algoritmo $brute$ que usará uma pesquisa de força bruta, métrica $manhattan$ que tem por objetivo ser usada para o cálculo da distância, $n\_neighbors$ = 1 que define que o melhor número de vizinhos é igual a 1 e o weights encontrado foi o $uniform$ que coloca todos os pontos em cada vizinhança ponderados igualmente.

Por outro lado, o SVM (Support Vector Machine) ou máquina de vetores de suporte, é uma generalização de um classificador simples e intuitivo chamado classificador de margem máxima, que consiste em encontrar um hiperplano que separa o conjunto de dados. Com o GridSearch fora encontrado os seguintes hiperparâmetros kernel RBF; C igual 10, comum a todos os kernels e gamma igual a 1, que é o coeficiente do kernel.

Além dos métodos já citados utilizamos o Multi Layer Perceptron(MLP), que simula uma rede neural artificial que tem como base o funcionamento do sistema nervoso central humano e tem por objetivo agrupar, classificar e até mesmo predizer. Tem-se como unidade fundamental o perceptron que possui como primeira etapa o funcionamento através do ajuste de pesos existente e os calcula encontrando uma saída, como etapa posterior é calculado o erro entre o valor encontrado e o valor real, com base nisso a medida que o erro é propagado faz-se um processo de otimização para minimizar os erros e os pesos serem atualizados. Quando possuímos vários perceptrons conectados temos uma rede neural artificial, a qual possui camadas de entrada, intermediárias e de saída, além de que no processo de treinamento que a rede neural atualiza os pesos e aprende as características dos dados.

Por fim, fora usado a classificação gaussiana tem como base a classificação probabilística, onde as suas previsões assume a forma de probabilidade de classe[5]. Condicionando uma distribuição a priori temos para a Classificação Gaussiana a seguinte formulação:

$$p(y_* \,|\mathcal{D}, \psi, \mathbf{x}_*) \;= \int p(y_*|f_*)p(f_*|\mathcal{D}, \psi, \mathbf{x}_*)df_*, \tag{4}$$

no qual p é a probabilidade de sucesso-p(y=1|**x**)- que está relacionada com uma função latente irrestrita f(**x**) que é mapeada para um intervalo unitário por uma transformação sigmoidal.

Para a etapa de hiperparametrização foi usado dois kernel's. O primeiro kernel utilizado foi o Radial Basis Function (RBF). Esse kernel é do tipo estacionário, ou seja, depende apenas da distância de dois dados não dependendo de seus valores absolutos, além do mais são invariantes. Matematicamente esse kernel é dado por:

$$\kappa\left(\mathbf{x}_i, \mathbf{x}_j\right) = exp\left(-\frac{d(\mathbf{x}_i, \mathbf{x}_j)^2}{2l^2}\right), \tag{5}$$

no qual $d(\cdot,\cdot)$ é a distância euclidiana entre os dados de teste (x_i) e treinamento (x_j) e l é o parâmetro de escala

da hiperparametrização. Por fim, o outro kernel utilizado fora o rational quadratic que pode ser visto como uma mistura de kernel RBF, mas com diferentes escalas de comprimento. Matematicamente esse kernel é dado por:

$$\kappa\left(\mathbf{x}_i, \mathbf{x}_j\right) = \left(1 + \frac{d(\mathbf{x}_i, \mathbf{x}_j)^2}{2\alpha l^2}\right)^{-\alpha}, \tag{6}$$

tendo que α e l tem que ser maior que 0, podendo ser visto como uma escala de mistura da raiz da função exponencial de covariância.

Para a análise quantitativa dos modelos fora ausado a Matriz de Confusão, Acurácia, Precisão, Revocação e F1-score, que são aplicáveis tanto a classificação binária como multivariada. A Matriz de confusão tem por objetivo apresentar a performace do modelo de forma visual, mostrando os dados que foram classificados corretamente e incorretamente e frequentemente é usada para analisar se o modelo está favorecendo alguma classe.

Em contrapartida, a acurácia tem por objetivo verificar a quantidade de dados classificados corretamente na matriz de confusão. A acurácia é definida por:

$$Acurácia = \frac{n_c}{n_t}, \tag{7}$$

no qual n_c número de fácies classificados corretamente e n_t é o número total de classes.

Por outro lado, a precisão é usada para avaliação de modelos e costumeiramente é usada com Revocação e F1-measure, seu objetivo é verificar os valores de verdadeiro positivo com os erros. Ademais, analisa erros do e sua métrica é bastante eficaz pois nos informa que quanto maior é a precisão maior e mais apurado é o modelo. A precisão é matematicamente escrita por

$$Precisao = \frac{VP}{VP\ +\ FP}. \tag{8}$$

Em adição temos a revocação que é usada conjuntamente com a precisão e quanto maior a revocação mais apurado é o modelo tem por objetivo verificar erros negativos, indicando quantas amostras o modelo conseguiu classificar corretamente.A revocação pode ser matematicamente escrita por

$$Revocacao = \frac{VP}{VP\ +\ FN}. \tag{9}$$

O F1-Score indica uma visão geral da qualidade do modelo, une precisão e revocação em uma única medida e quanto maior F1-score mais apurado, ademais é aplicado tanto a classificação binária quando a multiclasse. A sua desvantagem se dá quando o modelo está enviesado assim influenciando no F1-score. Matematicamente o F-score é dado por:

$$F1 - score = \frac{2\ *\ precisão\ *\ revocacao}{precisao + revocacao}. \tag{10}$$

Importante mencionar que estas métricas não podem ser analisadas separadamente [6]. Além destas citadas acima, existem outras tais como: ROC Curve e AUC curve. Tanto o “area under the curve (AUC)” quanto o “receiver operating characteristic curve (ROC)” é um gráfico que mostra o desempenho de um classificador binário em função de seu limite de corte (cuttoff)[4].

Resultados e Discussão

Após todas as etapas que concernem a aplicação das técnicas de aprendizagem de máquinas pode-se

verificar na Tabela 2 as técnicas que tiveram melhor desempenho, MLP, KNN e Gaussianas (com kernel RBF e Rotational Hyperparaemters). As pricipais métricas de performaces utilizadas foram: acurácia de fáceis, precisão, recall-revocação e f1-score. Na Tabela 2 podemos observar que estes métodos possuem alta precisão na verificação das camadas heterogêneas. Usando o GridSearch para encontrar os melhores hiperparâmetros, buscou-se melhorar tais resultados para os mesmos métodos de classificação da Tabela. Entretanto, neste caso nota-se que os métodos KNN, SVM, MLP melhoram sua precisão ao contrário dos outros métodos que não conseguiram aumentar sua precisão, acurácia, revocação e o F1-score. Com esses resultados obtidos foram selecionados os melhores métodos sem hiperparametrização - MLP, KNN e Gaussiana - e todos os métodos hiperparametrizados.

Tabela 2: Valores das métricas de avaliação dos modelos de classificação sem hiperparametrização e com hiperparametrização -KNN, MLP, Logistic Regression, SVM e Classificação Gaussiana.

Métricas	SVM Dado Teste	SVM Hiperparametrizado Dado teste	MLP Dado teste	MLP Hiper-parametrizado Dado Teste	Logistic Regression Dado Teste	Logistic Regression Hiperpara-metrizado Dado Teste	KNN Dado Teste	KNN Hiperpara-metrizado Dado Teste	Gaussiana Dado Teste	Gaussiana Hiperparametrizado RBF Dado Teste	Gaussiana Hiperparametrizado Rational Quadratic Dado Teste
Acurácia Classificação Fácies	0.58	0.71	0.61	0.71	0.55	0.55	0.69	0.73	0.64	0.56	0.56
Precisão Classificação Fáceis	0.60	0.72	0.62	0.72	0.55	0.55	0.70	0.74	0.66	0.57	0.57
Recall Classificação Fáceis	0.59	0.72	0.71	0.71	0.56	0.56	0.70	0.74	0.65	0.57	0.57
F1 Classificação Fáceis	0.56	0.71	0.61	0.71	0.54	0.54	0.69	0.74	0.64	0.55	0.55
Acurácia Classificação de Fáceis Adjacentes	0.92	0.91	0.92	0.91	0.90	0.90	0.90	0.92	0.90	0.91	0.89
Precisão Classificação de Fáceis Adjacentes	0.93	0.92	0.93	0.92	0.91	0.91	0.91	0.92	0.91	0.90	0.91
Recall Classificação de Fáceis Adjacentes	0.93	0.92	0.92	0.92	0.90	0.90	0.91	0.92	0.91	0.90	0.90
F1 Classificaçãode Fáceis Adjacentes	0.93	0.92	0.92	0.92	0.90	0.90	0.91	0.92	0.91	0.90	0.90

Nota-se na Tabela 3 que os melhores métodos são: MLP hiperparametrizado, Logistic Regression e Gaussiana não hiperparametrizados, Gaussiana com kernels Rational e RBF parametrizados. Ainda em relação a Tabela 3, não devemos olhar somente para acurácia mas essa acompanhada da precisão, vendo que essas possuem 0.65, 0.66, 0.62, 0.65 e 0.61 respectivamente.

Tabela 3: Resultados das métricas de avaliação dos modelos de classificação sem hiperparametrização e com hiperparametrização no poço cego.

Métricas	SVM Hiperpara-metrizado Dado Cego	MLP Não Hiper-parametrizada Dado Cego	MLP Hiperpara-metrizado Dado Cego	Logistic Regression Hiperpara-metrizado Dado Cego	KNN Hiperpara-metrizado Dado Cego	KNN Não Hiper-parametrizado Dado Cego	Gaussiana Não Hiper-parametrizado Dado Cego	Gaussiana RBF Dado Cego	Gaussiana Rational Dado Cego
Acurácia Classificação Fáceis	0.41	0.46	0.52	0.52	0.39	0.43	0.52	0.48	0.53
Precisão Classificação Fáceis	0.43	0.52	0.65	0.66	0.43	0.50	0.62	0.61	0.65
Recall Classificação Fáceis	0.41	0.46	0.52	0.53	0.39	0.43	0.52	0.48	0.53
F1 Classificação Fáceis	0.39	0.46	0.53	0.53	0.39	0.43	0.52	0.47	0.52
Acurácia Classificação de Fáceis adjacentes	0.87	0.95	0.95	0.93	0.88	0.90	0.95	0.95	0.95
Precisão Classificação de Fáceis adjacentes	0.88	0.95	0.95	0.96	0.90	0.92	0.95	0.95	0.95
Recall Classificação de Fáceis adjacentes	0.87	0.95	0.95	0.94	0.88	0.90	0.95	0.95	0.95
F1 Classificação de Fáceis adjacentes	0.87	0.95	0.95	0.94	0.88	0.91	0.95	0.95	0.95

A Tabela 2 nos dá a visão quantitativa dos métodos sem hiperparametrização e com hiperparametrização nos dados de teste, já a Tabela 3 nos permite fazer a análise de como o modelo generalizou. Além disso, foi feita uma análise qualitativa de como tais métodos generalizaram, por meio da Figura 2 conseguimos verificar a partir dos círculos feitos no gráfico as áreas onde o método conseguiu classificar com eficiência.

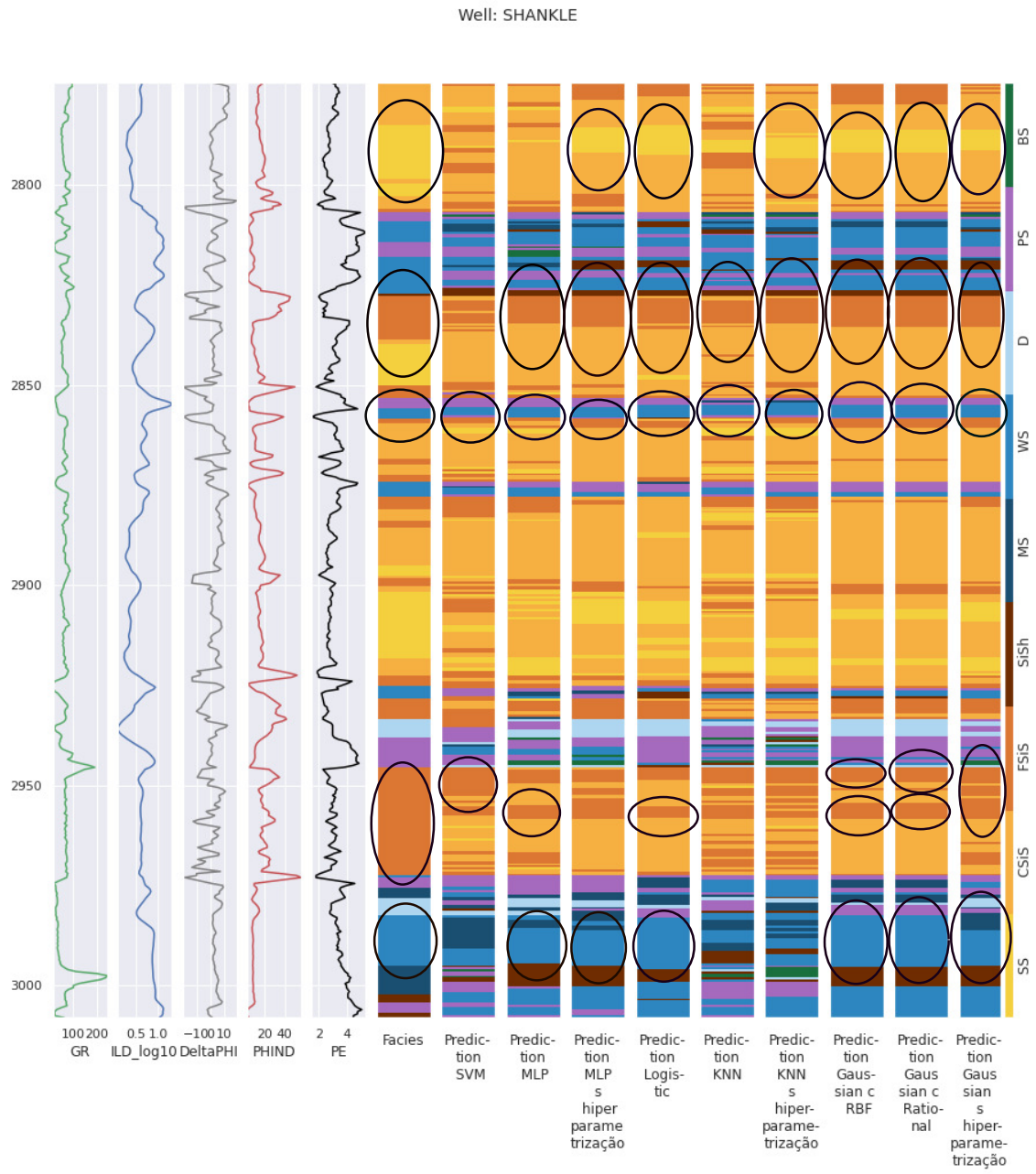

Figura 2: Mapa de fáceis classificadas corretamente denotados nas regiões destacadas em preto. Os método com piores performaces foram o SVM e o k-NN.

Conclusões

No presente trabalho, foram aplicado métodos de apredizagem de máquinas supervisionada paramétricos e não-paramétricos na classificação de fácies litológicas a partir de dados de poços. Dos métodos aplicados, destaca-se o método MLP que simula uma rede neural de forma paramétrica e não paramétrica. Também observou-se que a MLP paramétrica, uma rede neural simples, possui uma excelente precisão, uma boa acurácia para o presente modelo, assim como a regressão logística. Para o classificador Gaussiano não hiperparametrizado e hiperparametrizado com kernel RBF e Rational também apresentaram um bom redimento na classificação.

Em trabalhos futuros buscaremos otimizar estes modelos assim como buscar outros modelos. Também buscaremos modelos que possuam alto poder de generalização. Para isso será feito uma comparação entre os métodos de classificação em Deep Learning e Machine Learning, ou métodos ensembles.

Agradecimentos

Os agradecimentos são a Faculdade de Geofísica da UFPA, ao Programa de Pós-Graduação em Geofísica(CPGF/UFPA), ao Programa Institucional de Bolsas de Iniciação Científica-PIBIC/CNPq, pelo apoio ao desenvolvimento deste trabalho e pelo suporte dado devido a bolsa de iniciação científica, termo de cooperação entre PIBIC/CNPq/UFPA e pelo suporte dado devido ao Laboratório de Petrofísica e Física de Rochas Prof. Dr. Om Prakash Verma, que proporciona a utilização de computadores de alto desempenho.

Referências

[1] Martin K. Dubois, Geoffrey C. Bohling, and Swapan Chakrabarti. Comparison of four approaches to a rock facies classification problem. *Computers Geosciences*, 33(5):599–617, 2007.

[2] Y Zee Ma and Xu Zhang. *Quantitative geosciences: Data analytics, geostatistics, reservoir characterization and modeling*. Springer, 2019.

[3] S. Raschka. *Python Machine Learning*. Packt Publishing, Birmingham, UK, 1 edition, 2015.

[4] Tilo Strutz. *Data Fitting and Uncertainty: A Practical Introduction to Weighted Least Squares and Beyond*. Springer Vieweg, 2nd edition, 2015.

[5] C. K. Williams and C. E. Rasmussen. *Gaussian processes for machine learning*, volume 2. MIT Press, Massachusetts, USA, 2006.

[6] Eric R Ziegel. The elements of statistical learning, 2003.