COMPARAÇÃO DE MÉTODOS DE CLASSIFICAÇÃO PARA ANÁLISE DE DADOS PETROGRÁFICOS

Camila M. Saporetti
L. Goliatt

camilasaporetti@ice.ufjf.br
leonardo.goliatt@ufjf.edu.br
UFJF - Faculdade de Engenharia
Rua José Lourenço Kelmer, s/n, Campus Universitário, São Pedro, 36036-330, Juiz de Fora, MG

Leonardo C. Oliveira

leogeo.oliveira@petrobras.com.br
Petrobras
Avenida República do Chile, 65, Centro, 20031-912, Rio de Janeiro, RJ

E. Pereira

egberto@uerj.br
UERJ - Faculdade de Geologia
Rua São Francisco Xavier 524, 20559-900, Rio de Janeiro, RJ

Abstract. *This work aimed evaluate a methodology to measure different classification methods to predict the petrofacies of new thin-sections, which are removed by the same sedimentary basin. The data from boreholes drilled in the Parana Basin was processed in order to validate the proposed methodology. The presented methodology, besides evaluating the performance of various techniques of computational intelligence, is an alternative to assist the geologist/expert in the petrofacies determination and characterization, helping to reduce the manual effort involved in individualization process.*

Keywords: *Petrofacies, Computational Intelligence, Classification, Cross-Validation*

1 INTRODUÇÃO

A definição da distribuição de heterogeneidades de reservatórios de hidrocarbonetos são primordiais para exploração e otimização da produção de campos de petróleo. Para uma rocha ser julgada como um bom reservatório, ela deve possuir as seguintes características: uma extensão considerável, boa porosidade, uma apreciável permeabilidade e eficiência de recuperação de hidrocarbonetos. Estes atributos são ditos petrofísicos, e indicam o resultado de toda a história geológica dos sedimentos depositados e em particular das condições de sedimentação e dos fenômenos de diagênese, sendo de fundamental importância para definição da qualidade do reservatório.

O estudo da diagênese das rochas vem sendo incentivado pelas empresas petrolíferas, com intuito de entender a distribuição da porosidade em arenitos. O interesse vem do fato que estes arenitos podem se constituírem em rochas reservatórios de hidrocarbonetos. No decorrer do processo de diagênese, minerais podem precipitar-se como cimento nos poros da rocha, o que resulta na diminuição de sua porosidade e da permeabilidade, prejudicando seu potencial como reservatório [Maraschin e Mizusaki, 2008].

As petrofácies sedimentares são um conjunto de características petrográficas, como por exemplo, cor, granulometria, estrutura sedimentar, geometria deposicional, presença de fósseis, paleocorrente, entre outras que individualizam um grupo de rochas, e sua determinação permite a inferência da heterogeneidade do reservatório [Fávera, 2001]. A determinação de petrofácies sedimentares pode ser realizada através da contagem em microscópio petrográfico de luz transmitida, a fim de avaliar as heterogeneidades presentes no reservatório estudado. Este procedimento geralmente é longo, pois envolve o processo de amostragem, geração dos dados e posterior interpretações dos mesmos. Devido à grande quantidade de dados, nem toda a informação é utilizada com o método manual. Consequentemente, torna-se interessante o uso de análises automáticas a partir de ferramentas computacionais.

Nesse contexto, métodos de classificação aparecem como um mecanismo útil para auxiliar na identificação de petrofácies. Ela consiste na descoberta de regras de previsão para auxílio no planejamento e tomada de decisões. Utiliza-se esta técnica quando tem-se grandes quantidades de registros (amostras) em um banco de dados, que possuem diversos atributos, e tem-se como objetivo extrair desse banco de dados algum conhecimento relevante com capacidade preditiva. Nos últimos anos vários trabalhos têm usado técnicas baseadas em inteligência computacional para assistir na solução de problemas de caracterização de reservatórios. Maraschin e Mizusaki [2008] desenvolveram um método alternativo para prever permeabilidade, que se baseia na integração entre teoria wavelet e Redes Neurais Artificiais e aplicaram em dados de perfis de poços. O trabalho de da Silva et al. [2015] compara o desempenho de diversas técnicas de inteligência computacional para a tarefa de classificação da permeabilidade em um conjunto de dados petrofísicos com 78 amostras de rochas de seis reservatórios carbonáticos diferentes. Silva et al. [2015] realizaram a classificação petrográfica de rochas carbonato-siliciclástica usando um algoritmo de retropropagação de rede neural suportado por informações mineralógica e textural a partir de um conjunto de dados coletados da Bacia Provence do Sul, no sudoeste da França. Gholami et al. [2014] empregaram Regressão de Vetores Relevantes (Relevance Vector Regression, RVR) e Máquinas de Vetores Suporte (Support Vector Machines, SVM) combinados com Algoritmos Genéticos na previsão da

permeabilidade em dados de perfis retirados de três poços localizados em um reservatório carbonático na parte sul do Irã. Iturraran-Viveros e Parra [2014] aplicaram modelos de Redes Neurais Artificiais para prever a permeabilidade e a porosidade em dados de perfis de poços de um aquífero no sudeste da Flórida e a atenuação intrínseca de um reservatório de óleo de areia xistosa no nordeste do Texas. Olatunji [2011] usou dois tipos de sistemas lógica Fuzzy como uma nova abordagem para prever a permeabilidade de perfis de poços para a caracterização de um reservatório e Al-Anazi e Gates [2010] desenvolveram um modelo de previsão de porosidade baseado em Regressão de Vetores Suporte (Support Vector Regression, SVR) para um reservatório heterogêneo de arenitos.

Este trabalho tem o intuito de analisar uma metodologia para avaliar distintas técnicas baseadas em inteligência computacional para a previsão da classificação petrográfica de dados provenientes de diferentes intervalos deposicionais da Bacia Sedimentar do Paraná. Devido a natureza dos dados, são empregadas técnicas de balanceamento para melhorar o desempenho dos métodos que são comparados através de métricas de classificação e análise de significância.

2 MATERIAIS E MÉTODOS

2.1 Dados petrográficos

Os dados analisados são uma composição de 104 lâminas petrográficas obtidas a partir de amostras coletadas em 6 furos de sondagem pertencentes ao Tibagi (Paraná) e ao Dom Aquino (Mato Grosso do Sul) da Bacia Sedimentar do Paraná. A Bacia do Paraná é uma bacia que se localiza em uma região tectonicamente estável, que possui uma área de 1600000 km^2 com partes localizadas no Brasil, Argentina, Paraguai e no norte do Uruguai, como mostrado na Figura 1.

A análise quantitativa das amostras foi feita por meio da contagem em microscópio petrográfico de luz transmitida de 300 pontos em cada lâmina petrográfica com espaçamento de 0.3 mm. Nas amostras da Bacia do Paraná, foram contabilizadas, para cada lâmina, as porcentagens de 11 constituintes: quartzo, feldspato, glauconita, bioclasto, crescimento secundário de quartzo, caulinita, pirita, siderita, cimento carbonático, porosidade intergranular e porosidade intragranular.

De acordo com os resultados obtidos com a análise quantitativa dos dados petrográficos foi possível individualizar os arenitos (pela classificação manual) em 9 petrofácies distintas: PT-1, PT-2, PT-3,PT-4, I-1 e I-2 [Oliveira, 2009] e PG-1, PG-2 e PG-3 [Brazil, 2004]. A petrofácies I-1, PG-3 e PT-2 possuem somente uma amostra cada. As petrofácies PT-3, PG-2, PT-4, I-1, PT-1, PG-1 contém 2, 3, 5, 7, 28 e 56 amostras, respectivamente. Nota-se que as petrofácies PT-1 e PG-1 contêm 81% das amostras, ocasionado em uma base de dados desbalanceada.

A Tabela 1 mostra um exemplo de especificação dos dados. Cada coluna da tabela é uma amostra, que tem uma identificação (ex: PPG5-1), a profundidade em que a amostra foi retirada e as porcentagens dos atributos.

2.2 Método Computacional

A metodologia foi dividida em dois passos: (I) Realização de uma busca exaustiva com validação cruzada para encontrar os parâmetros ótimos para os classificadores e (II) emprego de métricas

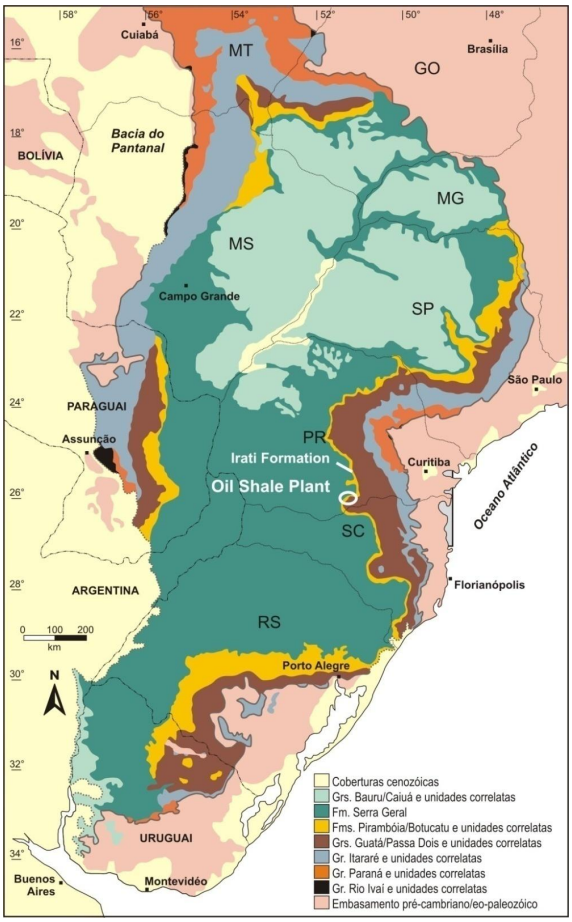


Figura 1: Localização da Bacia do Paraná (extraído de [França e Oliveira, 2010])

combinados com testes de significância para avaliar e comparar o desempenho dos métodos de classificação.

Aplicou-se métodos de classificação que utilizam diferentes abordagens, como K-Nearest Neighbors (KNN), Decision Trees (DT), Linear Support Vector Machine (LinSVM), RBF Support Vector Machine (RBFSVM) e Multi-Layer Perceptron (MLP). O desempenho desses métodos dependem do ajuste de parâmetros, e para auxiliar neste ajuste empregamos um procedimento de busca exaustiva com validação cruzada. A validação cruzada é uma técnica para avaliar a capacidade de generalização de um modelo, a partir de um conjunto de dados. Esta técnica é amplamente empregada em problemas onde o objetivo da modelagem é a predição. As técnicas de validação cruzada particionam o conjunto de dados em subconjuntos reciprocamente únicos, e depois, utiliza-se alguns destes subconjuntos para a estimação dos parâmetros do modelo (dados ou amostras de treinamento) e o restante dos subconjuntos (dados ou amostras de teste) são empregados na validação do modelo. O grid search é a busca exaustiva pelos melhores parâmetros a partir da análise dos resultados obtidos com a execução do algoritmo para um intervalo de parâmetros e pode ser usado em conjunto com a validação cruzada, como foi realizado neste trabalho. A faixa de variação dos parâmetros de cada modelo no procedimento de grid search com validação cruzada pode ser visto

Tabela 1: Exemplo das amostras retiradas do Poço PPG-5

Profundidade(m)	21, 5	46, 4	47, 1	55, 3
Lâmina delgada	PPG5-1	PPG5-2	PPG5-3	PPG5-4
Petrofácies	PT-1	PT-1	PT-3	PT-1
Quartzo	40,0	34,3	38,3	41,6
Feldspato	6,0	5,6	1,3	1,7
Muscovita	19,0	23,6	7,3	18,0
Opaco	0,6	1,3	1,0	3,4
Turmalina	0,0	0,0	0,0	0,0
Zircão	0,3	1,0	0,0	0,4
Rutilo	0,0	0,0	0,3	0,0
Glauconita	0,0	0,0	0,0	0,3
Clorita	0,0	0,0	0,0	0,7
Pseudo Matriz	1,3	7,0	2,0	12,7
Litoclasto	0,0	0,0	0,0	0,0
Bioclasto	0,0	0,0	0,0	0,0
Cresc. Sec. Qtz	3,4	3,0	7,6	1,0
Caolinita	10,0	11,3	8,0	1,0
Ilita/Smectita	10,4	12,3	6,3	13,4
Pirita	0,0	0,0	0,0	0,0
Siderita	0,0	0,0	0,0	0,0
Cim. Carbonático	0,0	0,0	27,6	1,4
Cim. Silicoso	0,0	0,0	0,0	0,7
Cim. Ferruginoso	0,0	0,0	0,0	1,4
Por. Intergranular	9,0	0,6	0,3	1,0
Por. Intragranular	0,0	0,0	0,0	0,0

na Tabela 2.

O método KNN faz uso dos K vizinhos mais próximos para estimar a classe de um novo padrão C, o algoritmo KNN calcula os K-vizinhos [Andoni e Indik, 2006]. mais próximos a C e atribui a classificação que aparece com maior frequência dentre os seus K-vizinhos Os parâmetros desse método são o número de vizinhos e a função de ponderação utilizados na previsão podendo ser uniforme ou ou calculado pela distância. As Árvores de Decisão (DT) são ferramentas que podem ser usadas para dar ao agente a capacidade de aprender, bem como para tomar decisões. Elas tomam como entrada uma situação descrita por um conjunto de atributos e retorna uma decisão, que é o valor predito para o valor de entrada. A decisão é tomada pela execução de uma sequência de testes [Dumont, 2009]: cada nó interno da árvore corresponde a um teste do valor de uma das propriedades, e os ramos deste nó são identificados com os possíveis valores do teste. Cada nó folha da árvore especifica o valor de retorno se a folha for atingida. Nesse método o parâmetro avaliado é a profundidade máxima da árvore.

Máquinas de Vetores Suporte (SVM) [Shanmugamani et al., 2015] são algoritmos de aprendi-

zagem de máquina que realizam a combinação linear de atributos por meio de funções, denominadas funções kernel, para atribuir uma classe a uma determinada amostra. Pode-se utilizar diferentes tipos de funções kernel e diferentes parâmetros a variar de acordo com o kernel selecionado. O SVM é comumente formulado como um problema de otimização da seguinte forma,

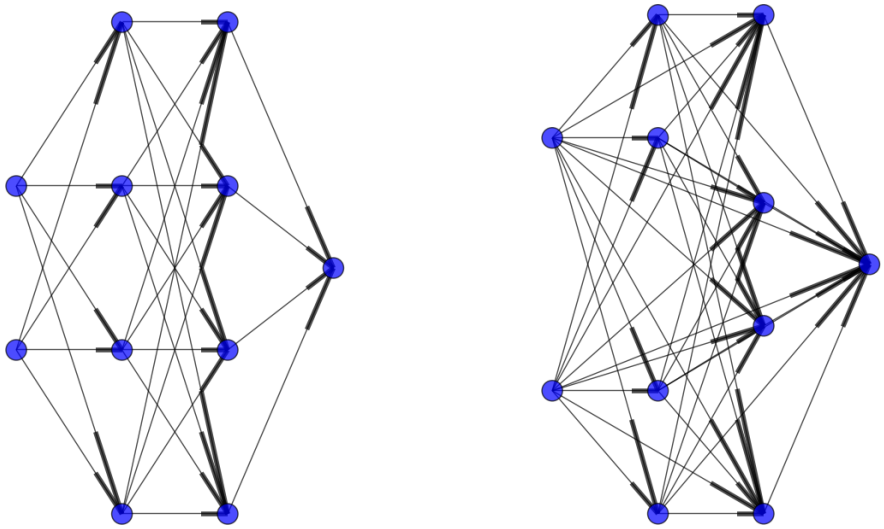
$$\begin{aligned} &\min_{w,b,\xi} \left(\frac{1}{2} w'w + C \sum_{i=1}^{n} \xi_i \right) \\ &\text{sujeito a} \quad y_i(w'\phi(z_i) + b) \geq 1 - \xi_i, \quad \xi_i \geq 0, \quad i = 1, 2, \ldots h, \end{aligned} \tag{1}$$

onde y_i são as classes da amostra de treinamento x_i, ϕ é usado para transformar os dados em espaço de alta dimensão, w representa os coeficients da função de decisão, a constante $C > 0$ é o erro de separação do hiperplano e ξ é usado para penalizar a função objetivo. O produto interno $\phi(z_i)'\phi(z_i)$ é substituído pelo kernel $K(z_i, z_j)$ que tem algumas propriedades especiais. Neste trabalho foram utilizados o kernel linear representado por $K(z_i, z_j) = z_i' z_j$ e o kernel de base radial $K(z_i, z_j) = e^{(-\gamma ||z_i - z_j||^2)}$, onde γ é um parâmetro da função de base radial. O desempenho dos métodos acima depende da escolha adequadas dos parâmetro C para o kernel linear e γ e C para o kernel RBF.

As redes Perceptron de Múltiplas Camadas, ou Multi-Layer Perceptron (MLP) [Haykin, 2001] têm sido aplicadas em variadas áreas, desempenhando funções de reconhecimento de padrões, controle e processamento de sinais, dentre outras. As redes de múltiplas camadas diferenciam-se das redes de camada simples pelo número de camadas intermediárias, camadas entre a camada de entrada e a de saída. Essa arquitetura tem uma ou mais camadas ocultas, que são constituídas por neurônios computacionais, também conhecidos como neurônios ocultos. A atividade dos neurônios ocultos é intervir entre a camada de entrada externa e a saída da rede de forma útil. Incluindo-se uma ou mais camadas ocultas, a rede torna-se apta de retirar estatísticas de ordem elevada. Este algoritmo tem como parâmetro o número de camadas ocultas e neurônios, o algoritmo, a conectividade e renomarlize. O número de camadas ocultas e neurônios é representado como [5,5], por exemplo, que indica 2 camadas ocultas com 5 neurônios cada. O algoritmo aborda os métodos usados para otimizar os pesos: l-bfgs é da família dos métodos Quase-Newton, sgd refere-se ao Método Descida Gradiente Estocástico, tnc trata a informação do gradiente no algoritmo de Newton truncado. A conectividade pode ser simplesmente conectada (mlgraph) ou totalmente conectada (tmlgraph). A Figura 2 exemplifica os tipos de conectividade. A renormalização se Verdadeira os dados são renormalizados.

Para avaliar os métodos descritos anteriormente, foram empregadas as estratégias de validação cruzada K-Fold (KF) e Stratified K-Fold (SKF) [Kohavi, 1995]. No KF a base de dados disponível contém N amostras, e é dividida em K subconjuntos, onde $K > 1$. Depois da partição da base de dados, os K-1 subconjuntos gerados são usados para treinamento e o conjunto restante é usado para teste, dessa maneira ao final do procedimento é medido o erro de validação. Esse processo é repetido K vezes, cada vez usando um conjunto de teste distinto em cada iteração. O intuito desse método é treinar da melhor forma possível a máquina para que ela possa generalizar sobre as futuras entradas. A Figura 3 mostra um esquema exemplificando o K-Fold.

No SKF, os subconjuntos são selecionados de modo que o valor médio da resposta seja apro-



(a) rede unidirecional simplesmente conectada

(b) rede unidirecional totalmente conectada

Figura 2: Esquema de ilustração das conectividades para uma rede neural 2-4-4-1 com duas entradas, duas camadas escondidas (4 neurônios em cada uma) e uma saída. Em (a) o esquema de conexões simples com os neurônios são conectados somente aos neurônios da camada anterior, e em (b) o esquema totalmente conectado onde os neurônios estão conectados a todos os antecessores.

Tabela 2: Parâmetros dos modelos utilizados no grid search com validação cruzada. O papel dos parâmetros pode ser explicado em [Pedregosa et al., 2011] e [Nissen, 2005].

Modelo	Parâmetros	Variação
KNN	no. de vizinhos	1,2,3,4,5,6,10
	pesos	uniforme, distância
DT	profundidade máxima	Nenhuma, 1,2,3,4,5,6,7,8,9,10,20,30,50
LinSVM	C	$10^{-2}, 10^{-1}, 10^{0}, 10^{1}, 10^{2}, 10^{3}, 10^{4}, 10^{5}, 10^{6}, 10^{7}, 10^{8}$
RBFSVM	C	$10^{-2}, 10^{-1}, 10^{0}, 10^{1}, 10^{2}, 10^{3}, 10^{4}, 10^{5}, 10^{6}, 10^{7}, 10^{8}$
	γ	$10^{-5}, 10^{-4}, 10^{-3}, 10^{-2}, 10^{-1}, 10^{0}, 10^{1}, 10^{2}, 10^{3}, 10^{4}$
MLP	no. de camadas ocultas e neurônios	[5],[5,5],[5,5,5],[5,5,5,5],[10],[10,10],[10,10,10],[50]
	algoritmo	tnc, l-bfgs, sgd, rprop, genetic
	conectividade	mlgraph, tmlgraph
	renormalização	Verdadeira,Falsa

ximadamente igual em todos os subconjuntos. Isto significa que cada um dos subconjuntos (trei-

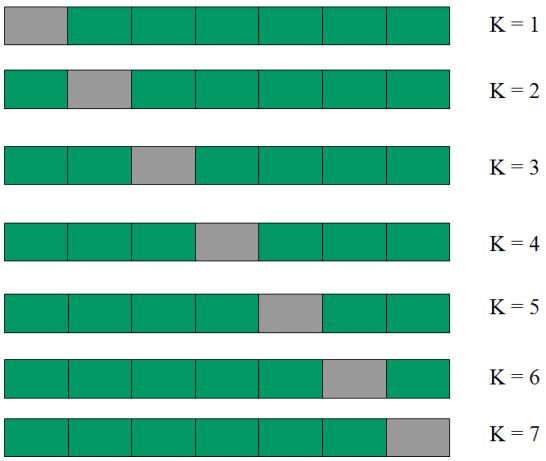


Figura 3: K-Fold - $K = 7$. Conjunto de treinamento, 6 amostras (quadros verdes) e conjunto de teste, 1 amostra (quadro cinza)

namento e teste) contém aproximadamente as mesmas proporções dos tipos de classe. Fez-se uso desse abordagem pelo fato de trabalharmos com dados desbalanceados. Na base de dados existem classes (petrofácies) que possuem poucas amostras classificadas, assim como tem-se petrofácies com muitas amostras classificadas.

A fim de tratar o problema do desbalanceamento também aplicou-se o a prática de oversampling, onde as amostras pertencentes as classes minoritárias foram duplicadas aletoriamente até obter uma base de dados balanceadas, que todas as classes tenham o mesmo número de amostras.

Realizou-se a avaliação dos métodos de classificação juntamente com os métodos de validação cruzada citados a partir das métricas acurácia (ACC), recall (RECALL) e F1 [Powers, 2007]. A acurácia mede a proporção das amostras serem classificadas corretamente, como pode ser visto na Eq. (2),

$$\text{ACC} = \frac{1}{N}\sum_{i=1}^{N} I(f(x_i) = y_i) \tag{2}$$

onde $f(x_i)$ é a classe predita pelo algoritmo de classificação e y_i é a classe original da amostra. Considera se $I(verdadeiro) = 1$ e $I(falso) = 0$. O recall pode ser referido como a taxa de verdadeiros positivos ou sensibilidade, como definido na Eq. (3)

$$\text{RECALL}(k) = \frac{TP}{TP + FN} \tag{3}$$

onde TP é taxa de verdadeiro positivo e FN é a taxa de falso negativo. A medida de desempenho F1 pode ser definida como valor positivo preditivo

$$\text{F1} = \frac{2TP}{(2TP + FP + FN)} \tag{4}$$

onde FP é a taxa de falsos positivos.

3 RESULTADOS E DISCUSSÃO

A metodologia proposta foi implementada na linguagem Python com módulos compilado na linguagem C, dentro do arcabouço do pacote Scikit-learn [Pedregosa et al., 2011] que integra uma gama de algoritmos de aprendizado de máquina para problemas supervisionados. Por intermédio das métricas citadas anteriormente realizou-se uma comparação das técnicas usadas na resolução do problema proposto.

A Tabela 3 apresenta a média e o desvio padrão da acurácia (ACC) , F1 e recall (RECALL) e a média do tempo de processamento para validação cruzada KF. Realizou-se uma análise estatística através do teste não-paramétrico de Wilcoxon e foi constatado que, usando o KF, os métodos apresentam desempenhos distintos. Observa-se que para a métrica ACC o RBFSVM obteve melhor desempenho, por esse motivo foi escolhido como base para o teste de Wilcoxon. Houve uma diferença de 9.2% para o DT, de 1.1% para o KNN, de 1.7% para o LinSVM e 3.1% para a MLP. Estas diferenças são consideradas estatisticamente significantes pelo teste de Wilcoxon. Resultados similares foram obtidos para as demais métricas.

Os resultados para a validação cruzada SKF são mostrados na Tabela 4. Nota-se que o RBFSVM foi o método com maior acurácia, F1 e recall. As diferenças na métrica ACC foram de 8% para o DT, 0% para o KNN, 1.5% para o LinSVM e 2.4% para o MLP, com o mesmo comportamento observado no RECALL e F1. Observa-se que o KNN que o obteve um desempenho comparável ao RBFSVM para as métricas ACC e RECALL.

O DT não obteve bons resultados comparado com os outros métodos, as árvores de decisão criam árvores tendenciosas se algumas classes possuem mais amostras que outras. O MLP0 é uma rede neuronal de múltiplas camadas, o faz com que a quantidade de conexões na arquitetura e na rede sejam grandes o que dificulta a convergência para a solução no processo de minimização do funcional do erro médio quadrático. O bom desempenho do KNN se deve ao fato que os valores de cada atributo foram normalizados, com intuito de que todos caiam num mesmo intervalo de variação, não havendo muita discrepância entre os valores dos diferentes atributos, que poderia influir tendenciosamente no cálculo da distância. Os SVMs obtiveram apreciável acurácia, por serem classificadores eficientes, na maioria dos problemas testados na literatura oos métodos apresentam bons resultados, e por utilizarem técnica de relaxamento minimizando o risco de overfitting, que é um problema que aparece em dados com grande dimensionalidade esparsos, como é o caso da base de dados utilizada. A diferença é o que o RBFSVM permite resolver problemas, originalmente, não linearmente separáveis, através do mapeamento para um espaço de maior dimensão.

As Figuras 4 e 5 apresentam os resultados dos dados sem o balanceamento. Para os classificadores DT, KNN, LinSVM, RBFSVM e MLP respectivamente. Nota-se que o uso da validação cruzada SKF resultou em melhores acurácia, F1 e recall para todos os métodos de classificação estudados. Esse resultado era esperado, uma vez que o SKF realiza uma melhor distribuição entre as classes das amostras entre os subconjuntos de treinamento e de teste, permitindo que estes conjuntos tenham a mesma proporção de amostras de cada classe (petrofácies).

O DT não obteve bons resultados comparado com os outros métodos, independente da abordagem de validação cruzada adotada. As árvores de decisão criam árvores tendenciosas se algumas

Tabela 3: Média e Desvio Padrão da Acurácia, F1 e recall para SKF. O TEMPO é o tempo médio de processamento em segundos do classificador para uma execução. Os melhores resultados estão em negrito enquanto * indica que a diferença observada não é estatisticamente significante com o respectivo melhor resultado. Um par de conjuntos de resultados são estatisticamente significantemente diferente quando o p-valor a partir do teste não-paramétrico de Wilcoxon é menor que 0.05.

MODELO	CV	ACC	F1	RECALL	TEMPO (s)
DT	KF	0.793 (0.020)	0.776 (0.027)	0.793 (0.020)	**0.001**
KNN	KF	0.863 (0.014)	0.834 (0.016)	0.863 (0.014)	0.004
LinSVM	KF	0.858 (0.014)	0.845 (0.011)	0.858 (0.014)	0.018
MLP	KF	0.846 (0.021)	0.836 (0.021)	0.846 (0.021)	1.394
RBFSVM	KF	**0.873 (0.014)**	**0.859 (0.019)**	**0.873 (0.014)**	0.023

Tabela 4: Média e Desvio Padrão da Acurácia, F1 e recall para SKF. O TEMPO é o tempo médio de processamento em segundos do classificador para uma execução. Os melhores resultados estão em negrito enquanto * indica que a diferença observada não é estatisticamente significante com o respectivo melhor resultado. Um par de conjuntos de resultados são estatisticamente significantemente diferente quando o p-valor a partir do teste não-paramétrico de Wilcoxon é menor que 0.05.

MODELO	CV	ACC	F1	RECALL	TEMPO (s)
DT	SKF	0.796 (0.021)	0.778 (0.023)	0.796 (0.021)	0.002
KNN	SKF	**0.877*(0.009)**	0.851 (0.009)	**0.877* (0.009)**	0.004
LinSVM	SKF	0.862 (0.014)	0.849 (0.012)	0.862 (0.014)	0.020
MLP	SKF	0.853 (0.021)	0.846 (0.020)	0.853 (0.021)	1.632
RBFSVM	SKF	**0.877 (0.010)**	**0.864(0.014)**	**0.877 (0.010)**	0.025

classes possuem mais amostras que outras. A base de dados em questão está desbalanceada, e o uso do SKF não foi suficiente para que o classificador gerasse bons resultados em comparação com os outros métodos. A partir dessa observação, os dados foram balanceados e empregou-se o validação cruzada SKF, a fim de avaliar o impacto do tratamento dos dados no desempenho dos classificadores. Os resultados deste segundo procedimento pode ser vistos na Tabela 5 e na Figura 6.

A Tabela 5 apresenta a média da acurácia, F1 e recall e o tempo de processamento. Realizou-se uma análise estatística através do Teste não-paramétrico de Wilcoxon e foi constatado que os métodos são estatisticamente diferentes com exceção do KNN, que obteve de acordo com teste de Wilcoxon desempenho similar ao RBFSVM, método que foi tomado como referência para o teste de Wilcoxon.

Depois de realizado o balanceamento dos dados e após a análise dos resultados percebe-se que

Tabela 5: Média e Desvio Padrão da Acurácia, F1 e recall para SKF. O TEMPO é o tempo médio de processamento em segundos do classificador para uma execução. Os melhores resultados estão em negrito enquanto * indica que a diferença observada não é estatisticamente significante com o respectivo melhor resultado. Um par de conjuntos de resultados são estatisticamente significantemente diferente quando o p-valor a partir do teste não-paramétrico de Wilcoxon é menor que 0.05.

MODELO	CV	ACC	F1	RECALL	TEMPO (s)
DT	SKF	0.977 (0.002)	0.977 (0.002)	0.977 (0.002)	**0.004**
KNN	SKF	0.992 (0.001)*	0.992 (0.001)*	0.992 (0.001)*	0.007
LinSVM	SKF	0.985 (0.003)	0.985 (0.003)	0.985 (0.003)	0.045
MLP	SKF	0.986 (0.004)	0.986 (0.004)	0.986 (0.004)	2.975
RBFSVM	SKF	**0.995 (0.003)**	**0.995 (0.003)**	**0.995 (0.003)**	0.055

o método que ocasionou melhor acurácia, F1 e recall foi o RBFSVM, seguido do KNN. Ambos os métodos se mostraram úteis para a classificação de dados petrográficos, tanto desbalanceados como balanceados. Os classificadores MLP , LinSVM e DT apresentaram bons resultados e mostraram ser uma alternativa para bases de dados balanceadas. No entanto, nota-se com base nos resultados apresentados que o balanceamento dos dados foi um fator que proporcionou ganho de desempenho em todos os métodos.

4 CONCLUSÕES

No presente trabalho foram analisados os desempenhos de cinco classificadores (KNN, DT, LinSVM, RBFSVM e MLP) para a classificação de petrofácies sedimentares. Os resultados experimentais demonstraram que tal metodologia é eficaz para a tarefa proposta, com acurácia em torno de 86% dos casos de teste, com exceção do DT, nos dados desbalanceados. Já nos dados balanceados os classificadores apresentaram acurácia, F1 e recall em torno 98%, mostrando que, para o conjunto de dados analisados, o balanceamento influencia positivamente o desempenho dos classificadores. Os métodos RBFSVM e KNN foram os que mais se destacaram, 87.7% de acurácia nos dados desbalanceados e 99% de acurácia nos dados balanceados. Como o tempo médio desses dois processamentos foi próximo, os dois métodos mostraram-se eficientes na resolução do problema proposta. A ferramenta computacional desenvolvida permite obter resultados semelhantes aqueles obtidos pelo método convencional de individualização, atingindo uma boa precisão.

AGRADECIMENTOS

Os autores agradecem à Universidade Federal de Juiz de Fora (UFJF), ao Programa de Pós-Graduação em Modelagem Computacional (PPGMC), à CAPES (Procad 071/2013), à FAPEMIG (PCE-01337-15) pelo suporte dado e à Universidade do Estado do Rio de Janeiro (UERJ) por conceder acesso às lâminas utilizadas para análise e obtenção dos dados.

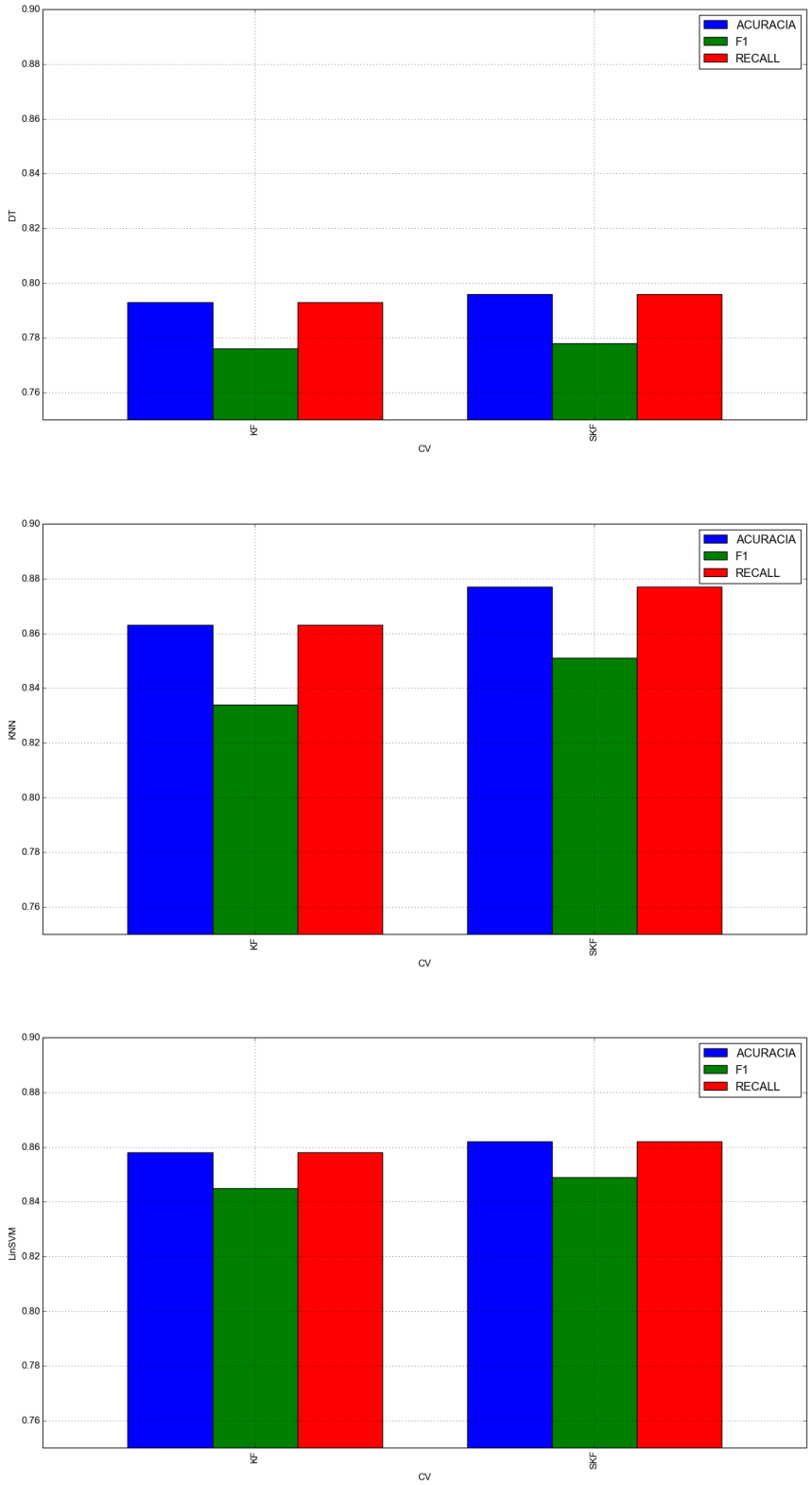


Figura 4: Dados Desbalanceados - DT, KNN e LinSVM (30 iterações, $k = 5$) - acurácia, F1 e recall.

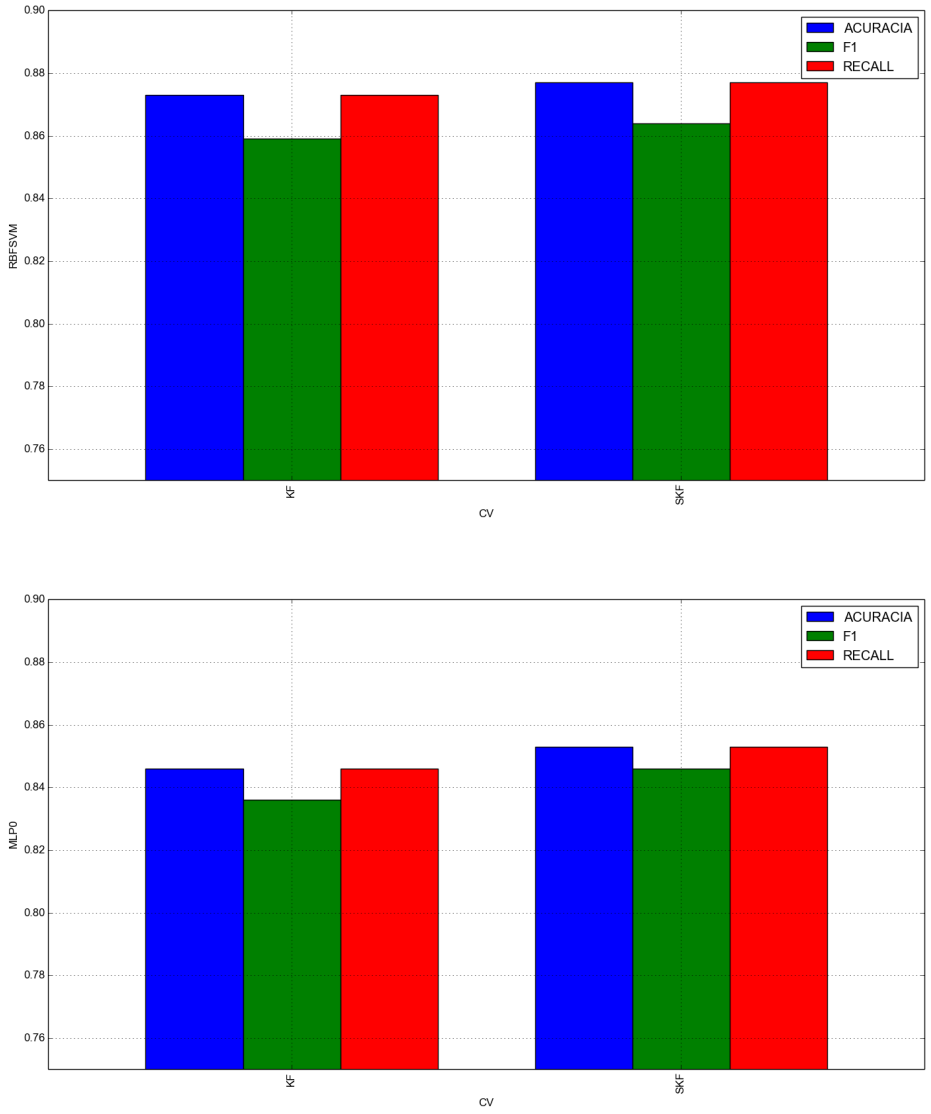


Figura 5: Dados Desbalanceados - RBFSVM e MLP (30 iterações, $k = 5$) - acurácia, F1 e recall.

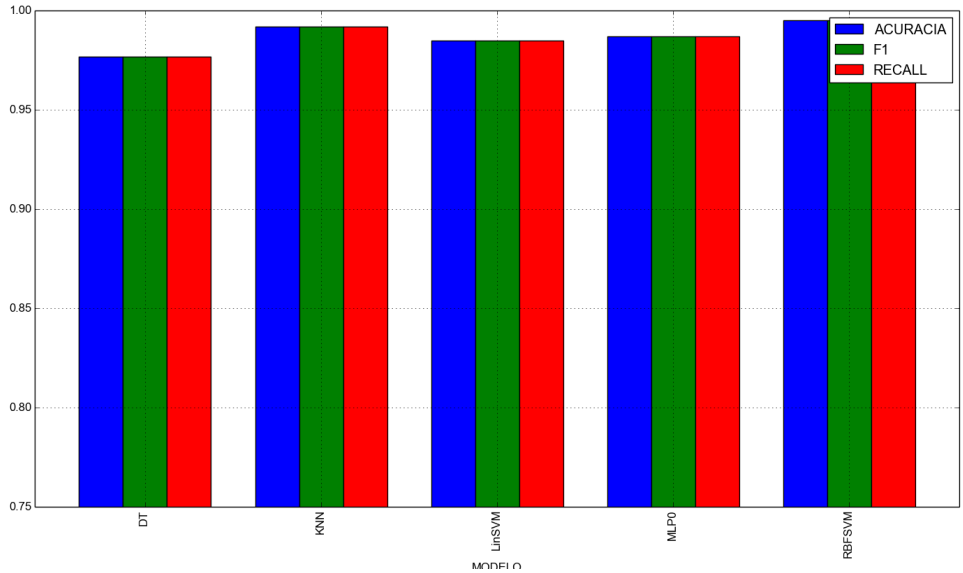


Figura 6: Dados Balanceados (30 iterações, $k = 5$) - acurácia, F1 e recall.

REFERÊNCIAS

A.F. Al-Anazi e I.D. Gates. Support vector regression for porosity prediction in a heterogeneous reservoir: A comparative study. *Computers and Geosciences*, 36(12):1494 – 1503, 2010. ISSN 0098-3004.

A. Andoni e P. Indik. Near-optimal hashing algorithms for approximate nearest neighbor in high dimensions. Foundations of Computer Science, 2006.

Fátima A. F. Brazil. Estratigrafia de sequências e processo diagenético: Exemplos dos arnitos marinho-rasos da formação ponta grossa, noroeste da bacia do paraná. Dissertação de Mestardo, Universidade Do Estado Do Rio De Janeiro, 2004.

Pablo Nascimento da Silva, Eduardo Corrêa Gonçalves, Edmilson Helton Rios, Asif Muhammad, Adam Moss, Tim Pritchard, Brent Glassborow, Alexandre Plastino, e Rodrigo Bagueira de Vasconcellos Azeredo. Automatic classification of carbonate rocks permeability from 1h {NMR} relaxation data. *Expert Systems with Applications*, 42(9):4299 – 4309, 2015. ISSN 0957-4174. URL http://www.sciencedirect.com/science/article/pii/S0957417415000494.

M. Dumont. Fast multi-class image annotacion with random subwindows and multiple output randomized trees. International Conference on Computer Vision Theory and Applications, 2009.

A. B. França e C. G. Oliveira. Bacia do paraná: Rochas e solos. Palestra Almerio Por Claudinei, 2010.

J. C. D. Fávera. *Fundamentos De Estratigrafia Moderna*. Eduerj, 1 edição, 2001.

Raoof Gholami, Ali Moradzadeh, Shahoo Maleki, Saman Amiri, e Javid Hanachi. Applications of artificial intelligence methods in prediction of permeability in hydrocarbon reservoirs. *Journal of Petroleum Science and Engineering*, 122:643 – 656, 2014. ISSN 0920-4105.

S. Haykin. *Redes Neurais*. Bookman, 2 edição, 2001.

Ursula Iturraran-Viveros e Jorge O. Parra. Artificial neural networks applied to estimate permeability, porosity and intrinsic attenuation using seismic attributes and well-log data. *Journal of Applied Geophysics*, 107:45 – 54, 2014. ISSN 0926-9851. URL http://www.sciencedirect.com/science/article/pii/S092698511400144X.

R. Kohavi. A study of cross-validation and bootstrap for accuracy estimation and model selection. 2:1137–1143, 1995.

Anderson José Maraschin e Ana Maria Mizusaki. Datação de processos diagenéticos em arenitos-reservatório de hidrocarbonetos: Uma revisão conceitual. 35(1):27–41, 2008.

Steffen Nissen. Neural networks made simple. *Software 2.0*, 2:14–19, 2005.

Sunday O. Olatunji. Modeling the permeability of carbonate reservoir using type-2 fuzzy logic systems. *Computers in Industry*, 62(2):147 – 163, 2011. ISSN 0166-3615. URL http://www.sciencedirect.com/science/article/pii/S0166361510001569. Fuzziness in Industry and Applications.

L. C. Oliveira. Estudos das relações entre o arcabouço estratigráfico e as alterações diagenéticas observadas na seção devoniana da bacia do paraná. Dissertação de Mestardo, Universidade Do Estado Do Rio De Janeiro, 2009.

F. Pedregosa, G. Varoquax, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, e E. Duchesnay Duchesnay. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830, 2011.

D. M. W. Powers. Evaluation: From precision, recall and f-factor to roc, informedness, markedness and correlation. Technical Report SIE-07-001, 2007.

Rajalingappaa Shanmugamani, Mohammad Sadique, e B. Ramamoorthy. Detection and classification of surface defects of gun barrels using computer vision and machine learning. *Measurement*, 60:222 – 230, 2015. ISSN 0263-2241. URL http://www.sciencedirect.com/science/article/pii/S0263224114004734.

Adrielle A. Silva, Irineu A. Lima Neto, Roseane M. Misságia, Marco A. Ceia, Abel G. Carrasquilla, e Nathaly L. Archilha. Artificial neural networks to support petrographic classification of carbonate-siliciclastic rocks using well logs and textural information. *Journal of Applied Geophysics*, 117:118 – 125, 2015. ISSN 0926-9851.