





Comparação entre Algoritmos de Aprendizado de Máquina para a Identificação de Floresta Tropical Sazonalmente Seca

Comparison between Machine Learning Algorithms for the Identification of Seasonally Dry Tropical Forest

Elisiane Alba¹ , Marta Laura de Souza Alexandre¹ , Juliana Marchesan² ,
Luciana Sandra Bastos de Souza¹ , Alan César Bezerra¹  & Emanuel Araújo Silva³ 

¹Universidade Federal Rural de Pernambuco, Unidade Acadêmica de Serra Talhada, Serra Talhada, Pernambuco, Brasil

²Secretaria da Agricultura, Pecuária e Desenvolvimento Rural, Departamento de Diagnóstico e Pesquisa Agropecuária, Centro Estadual de Diagnóstico e Pesquisa Florestal, Santa Maria, Rio Grande do Sul, Brasil.

³Universidade Federal Rural de Pernambuco, Departamento de Ciências Florestais, Dois Irmãos, Recife, Pernambuco, Brasil

E-mails: elisiane.alba@ufrpe.br; lauramarta012@gmail.com; juliana-marchesan@seapdr.rs.gov.br; sanddrbastos@yahoo.com.br; alan.bezerra@ufrpe.br; emanuel.ufrpe@gmail.com

Autor Correspondente: Elisiane Alba; elisiane.alba@ufrpe.br

Resumo

O objetivo do estudo foi avaliar o potencial dos algoritmos de aprendizado de máquina *k-Nearest Neighbor* (kNN), *Random Forest* (RF), *Support Vector Machine* (SVM) e *Artificial Neural Networks* (ANN) na identificação das fitofisionomias da Caatinga a partir de imagens OLI/Landsat 8. Para tanto, foi elaborado um arquivo vetorial de treinamento com as amostras dos diferentes padrões dos usos e cobertura da terra, a fim de orientar os algoritmos no processo de classificação. A validação das classificações foi obtida por meio da validação cruzada, do tipo *k-fold*, com 30 repetições, sendo avaliada a qualidade da classificação a partir dos valores expressos pelo coeficiente Kappa. Para verificar a existência de diferenças significativas entre os algoritmos foi aplicado o teste estatístico de Friedman e Nemenyi. O algoritmo RF apresentou os maiores valores para o coeficiente Kappa, expressando um valor médio de 0,9841. Por outro lado, a ANN demonstrou desempenho inferior aos demais, englobando um valor médio de 0,7551, ocasionado pela confusão espectral na identificação da classe nuvem/sombra com a classe água. Apesar de todos os algoritmos testados apresentarem bons resultados, o algoritmo RF diferiu significativamente dos demais, expressando resultados superiores quando aplicado à identificação de padrões espaciais na Caatinga. Conclui-se que o uso de algoritmos de aprendizagem de máquina é eficiente na identificação de fitofisionomias da Caatinga, com destaque para o RF, o qual englobou melhor a variação dos padrões espectrais dos usos, podendo ser utilizado para estudos posteriores relacionados com a Caatinga.

Palavras-chave: Inteligência artificial; Mapeamento da vegetação; OLI/Landsat 8

Abstract

The aim of the study was to evaluate the potential of the K-Nearest Neighbor (kNN), Random Forest (RF), Support Vector Machine (SVM) and Artificial Neural Networks (ANN) machine learning algorithms for the identification of Caatinga phytophysiological patterns from OLI/Landsat 8 images. For this purpose, a vector training file was elaborated with samples of the different patterns of land uses and cover, in order to guide the algorithms in the classification process. The validation of the classifications was obtained through the cross-validation, of the k-fold type, with 30 repetitions, being evaluated the quality of the classification from the values expressed by the Kappa coefficient. To verify the existence of significant differences between the algorithms, Friedman and Nemenyi's statistical test was applied. The RF algorithm presented the highest values for the Kappa coefficient, expressing an average value of 0.9841. On the other hand, ANN demonstrated inferior performance than the others, encompassing an average value of 0.7551, caused by spectral confusion in the identification of the cloud/shadow class with the water class. Although all algorithms tested present good results, the RF algorithm differed significantly from the other ones, expressing superior results when applied to the identification of spatial patterns in the Caatinga. It is concluded that the use of machine learning algorithms is efficient in the identification of Caatinga phytophysiological patterns, especially RF, which better encompassed the variation of spectral patterns of uses and can be used for further studies related to Caatinga.

Keywords: Artificial intelligence; Vegetation mapping; OLI/Landsat 8

Recebido em: 13/01/2021; Aprovado em: 15/02/2022

Anu. Inst. Geociênc., 2022;45:40758

DOI: https://doi.org/10.11137/1982-3908_45_40758 1



1 Introdução

O sensoriamento remoto tem sido uma importante ferramenta para a observação da dinâmica dos ecossistemas, permitindo análises periódicas e grandes extensões sobre a área de interesse. Um dos principais produtos do sensoriamento remoto consiste nas imagens da série Landsat, as quais são disponibilizadas gratuitamente aos usuários desde a década de 80. Com o desenvolvimento da área de ciência e computação, as técnicas de Inteligência Artificial e aprendizado de máquina (machine learning) ganharam destaque e aplicações nas mais diversas ciências, em especial, o uso combinado de dados de sensoriamento remoto. Nesse campo do conhecimento, é atribuído à máquina habilidades que simulam a inteligência humana, possibilitando a realização de funções que antes eram exclusivamente dos seres humanos (Fernandes 2003).

O aprendizado de máquina envolve a implementação de algoritmos, como o *Support Vector Machine* (SVM), *Random Forest* (RF), *K-Nearest Neighbor* (kNN) e *Artificial Neural Networks* (ANN), os quais podem aprender automaticamente a partir de uma amostra de dados disponibilizados. O algoritmo de aprendizagem de máquina SVM têm como objetivo a determinação de limites de decisão que produzam uma separação ótima entre classes por meio da minimização dos erros (Nascimento et al. 2009). O SVM define um hiperplano com base no intervalo máximo dos conjuntos de amostras de treinamento fornecidos e, em seguida, classifica os objetos em uma das classes de cobertura da terra identificada (Qian et al. 2015).

O algoritmo RF segue um objetivo diferente, uma vez que pode originar várias árvores de decisão por meio de conjuntos de atributos escolhidos de forma aleatória em relação ao conjunto original (Gaiad et al. 2017). Segundo Montañó et al. (2017), o RF compreende uma técnica de aprendizado de máquina que agrupa várias árvores de modelos treinadas a partir de um conjunto de dados para predição, de forma que o resultado é obtido por meio da consolidação dos resultados das árvores. Nesse sentido, Rodríguez e Chica (2012) descrevem que o RF aumenta a diversidade de árvores de decisão a partir de diferentes subconjuntos de dados criados usando um procedimento chamado ensacamento.

O algoritmo kNN representa uma técnica de aprendizado supervisionado que pode resolver um problema de classificação de amostra baseado na distância que existe entre um determinado exemplo de teste e amostras previamente classificadas (Alcantara, Ferreira & Santos 2018). Esse algoritmo é uma extensão do algoritmo 1-Vizinho Mais Próximo (1-NN). Por sua vez, as ANN são modelos computacionais inspirados no sistema nervoso de seres vivos, formado por um conjunto de unidades de

processamento caracterizadas por neurônios artificiais, interligados por inúmeras interconexões, denominadas de sinapses artificiais, permitindo a aquisição e manutenção do conhecimento (Soares & Teive 2015). Desse modo, o algoritmo ANN é robusto além de apresentar uma aprendizagem eficiente (Zheng et al. 2013).

Neste sentido, esses algoritmos podem apresentar ganhos quando aplicados à identificação dos diferentes usos e cobertura da terra, especialmente em regiões onde o comportamento espectral de alvos é similar, como é o caso da Caatinga, também denominada de Floresta Tropical Sazonalmente Seca (FTSS). As FTSS geralmente apresentam déficit hídrico, altas taxas de insolação e precipitação reduzida, concentrada em curtos períodos (Araújo et al. 2012). A fitofisionomia das FTSS é muito variada podendo-se encontrar áreas de vegetação arbustiva baixa e rala até florestas densas que podem atingir até cerca de 10m de altura (IBGE 2004).

No período de estiagem, a vegetação florestal perde suas folhas e tem sua resposta espectral confundida com solo exposto. Dessa forma, o mapeamento dessas coberturas em regiões semiáridas da Caatinga pelos métodos tradicionais, como por exemplo o classificador Maxver (Máxima Verossimilhança), envolve grandes erros e perda da qualidade das classificações.

Aliado a isso, a Caatinga compreende uma região com constantes alterações na paisagem devido a expansão das atividades agrícolas, perdendo gradualmente a cobertura nativa do bioma Caatinga (Gariglio et al. 2010). O município de Petrolina por sua vez, possui pressões antrópicas decorrentes da atividade agrícola, destacando-se a existência de forte pressão sobre as áreas de vegetação ciliar, vegetação primária, para a implantação de projetos agrícolas, assentamentos rurais, estimulando a degradação dos solos e vegetação (Taura et al. 2011). Essas alterações podem modificar as interações entre a biosfera e atmosfera, tendo como consequência a variação dos padrões climáticos em escala local e regional, bem como no ciclo hidrológico os quais podem resultar em alterações para as atividades e economia local. Estudos que monitoram as alterações na paisagem que têm influência climática têm sido realizados in loco e associação do uso de técnicas tradicionais de sensoriamento remoto. Entretanto, estudos relacionando a IA na observação dos padrões espectrais da Floresta Tropical Sazonalmente Seca permanecem inexistentes.

Nesse contexto, os algoritmos de aprendizado de máquina têm fornecido ganhos nos processos de regressão (Marchesan et al. 2020; Montañó et al. 2017), bem como na classificação de uso e cobertura da terra (Alba 2020; Gaiad et al. 2017; Qian et al. 2015), sendo um importante método a ser aplicado na observação das florestas secas.

O objetivo desse estudo foi avaliar o potencial dos algoritmos de aprendizado de máquina kNN, RF, SVM e ANN na identificação dos diferentes usos e cobertura da terra da Floresta Tropical Sazonalmente Seca a partir de imagens OLI/Landsat8.

2 Material e Métodos

2.1 Caracterização da Área de Estudo

A área de estudo correspondeu ao município de Petrolina, localizado no estado de Pernambuco, o qual possui uma área de 4.561,590 km². Em extensão territorial, Petrolina se destaca como um dos maiores municípios do estado de Pernambuco. Situa-se a 09° 23' 03" de latitude Sul e 40° 30' 22" de longitude Oeste (Figura 1). E têm suas atividades econômicas especialmente relacionadas à agricultura graças à disponibilidade hídrica do Rio São Francisco.

A área possui uma vegetação composta por Caatinga Hiperxerófila com trechos de Floresta Caducifólia. Possui diferentes tipos de solo destacando-se as seguintes classes: Argissolos Vermelho-Amarelos e Amarelos, ambos Eutróficos plínticos e não plínticos, abruptos ou não abruptos, concrecionários e não concrecionários.

Em menores proporções ocorrem áreas de Latossolos Vermelho-Amarelos Eutróficos, Neossolos Quartzarênicos Distróficos, Neossolos Litólicos Eutróficos, Planossolos Nátricos e Planossolos Solódicos (Silva et al. 2006).

O clima da região de acordo com a classificação de Koppen é Bsh, ou seja, semiárido. Seu clima é tropical semiárido seco e quente na parte norte e semiárido quente estépico na parte sul. Caracteriza-se pela escassez e irregularidade pluviométrica, com chuvas e forte evaporação em consequência das altas temperaturas no verão. O total anual médio de precipitação pluvial é da ordem de 560mm. As chuvas ocorrem de janeiro a abril devido ao deslocamento da Zona de Convergência Intertropical (ZCIT) em direção ao Hemisfério Sul, que influencia na convergência de umidade e da convecção local. Março e agosto são os meses com a maior e a menor precipitação, com totais médios de 136 mm e 5mm, respectivamente. Nos meses mais úmidos, a umidade relativa do ar varia, em média, entre 66% e 72%. Menores valores acontecem nos meses mais quentes, de setembro a novembro, quando a umidade atinge valores abaixo de 55%. A umidade relativa atinge os maiores valores em abril, que corresponde ao fim do período chuvoso. A temperatura do ar apresenta variações médias entre 24° C e 28° C, sendo julho o mês mais frio e novembro o mês mais quente do ano (SONDA 2021).

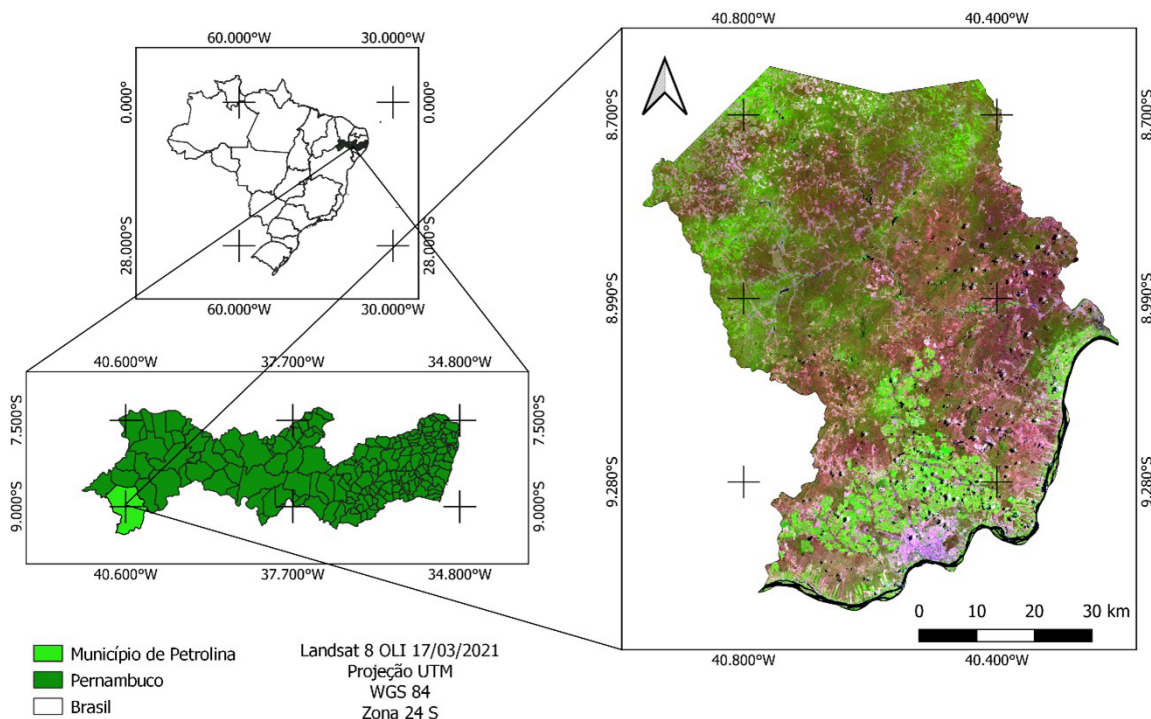


Figura 1 Mapa de localização da área de estudo

2.2 Dados Espaciais

Em relação aos dados espectrais, dispôs-se de uma imagem do sensor OLI (*Operational Land Imager*), abordo do satélite Landsat 8, da banda 1 até a banda 7. Estes dados foram obtidos do portal do Serviço Geológico Americano – USGS (*United States Geological Survey*), órbita-ponto: 217/66, datada de 24/07/2020, em que as imagens foram disponibilizadas ortorretificadas, com nível 1 de processamento.

Enquanto o dado vetorial, para a delimitação da área de estudo, foi disponibilizado pelo IBGE (Instituto Brasileiro de Geografia e Estatística) por meio da Malha Municipal do ano de 2019, no formato *shapefile*. Esse arquivo foi utilizado como máscara para o recorte das sete bandas espectrais.

2.3 Processamento dos Dados

Inicialmente, foi elaborado um arquivo vetorial de treinamento com mais de 300 amostras, a fim de orientar os algoritmos no processo de classificação. Assim, as amostras das classes de uso e cobertura da terra foram: água, vegetação arbórea (vegetação arbórea e arbustiva), áreas agrícolas (culturas agrícolas em desenvolvimento), solo exposto (afloramento rochoso e/ou solo sendo preparado para um novo ciclo de cultivo agrícola) e nuvem/sombra. As presentes classes foram delimitadas com base na necessidade local de conhecimento dos usos e cobertura da terra de forma a subsidiar ações de manejo e conservação do bioma.

A identificação das amostras de cada classe foi realizada com base na resposta espectral das fisionomias, utilizando a imagem OLI/Landsat 8 na sua composição RGB-654 (Falsa cor), juntamente com o *Plugin Quick Map Services*, disponível no Qgis. Desse modo, foram vetorizados polígonos para cada uma das classes temática descritas acima.

A classificação utilizada no presente estudo foi do tipo supervisionada, realizada por meio da técnica de aprendizado de máquina baseando-se nos algoritmos kNN, RF, SVM e ANN. Para cada algoritmo foi modelado a configuração ótima, de acordo com os parâmetros exigidos, conforme exposto na Tabela 1. Desse modo, foram testados diferentes parâmetros como o número de neurônios para o algoritmo ANN, tamanho da árvore para o RF, valor de custo e tipo de Kernel para o SVM, e por fim, o número de vizinhos ótimo para o classificador kNN.

Com o intuito de verificar o potencial de generalização dos algoritmos, foi aplicado a validação cruzada, uma vez que consiste em um método robusto e frequentemente aplicada na identificação do ajuste das

técnicas de aprendizado de máquina (Alba 2020), seja nas operações de classificação, como de regressão. Assim, a validação das classificações foi realizada por meio da validação cruzada, do tipo *k-fold*, com 30 repetições, sendo avaliado o coeficiente Kappa. No método de validação cruzada, para cada repetição, o conjunto de dados da amostragem é dividido em dois subconjuntos: treino e teste, sendo o coeficiente Kappa obtido por meio da análise das amostras de teste (não utilizadas para o treinamento dos algoritmos).

O Coeficiente Kappa, por sua vez, é utilizado para avaliar a qualidade da classificação, o qual segundo Moreira (2005), inclui em seu cálculo todos os elementos da matriz de erro, sendo expresso por valores entre 0 e 1, de modo que valores mais próximos de 1 correspondem a uma classificação mais precisa. Landis e Koch (1977) estabeleceram a qualidade das classificações segundo o intervalo de Kappa, desse modo, uma classificação com valor de Kappa superior a 0,60 é considerada como “muito boa”, sendo esse parâmetro utilizado no presente estudo.

Em posse dos valores de Kappa, obtido em cada repetição, utilizou-se o teste estatístico de Friedman (1940) e Nemenyi (1963) para verificar a existência de diferenças significativas entre os algoritmos ao nível de significância de 5%. Quando a hipótese nula é rejeitada (p -valor $< 0,05$), entende-se que há pelo menos um modelo (algoritmo) com desempenho significativamente diferente dos demais (Demsar 2006). Posteriormente, foi realizada uma análise comparativa, identificado o melhor algoritmos para a identificação dos usos em áreas de Caatinga.

A Figura 2 demonstra o fluxograma para o processamento da imagem, bem como o treinamento dos algoritmos de aprendizado de máquina englobados nesse estudo. Como produto, apresenta-se o mapeamento do uso e cobertura da terra para ambos os algoritmos em estudo.

A amostragem das classes de uso e cobertura da terra para o treinamento dos algoritmos foi realizada no Software Qgis versão 3.10.8. Posteriormente, o aprendizado dos algoritmos, bem como a espacialização das classes temáticas foram processadas no software R, versão 3.6.1 (R Development Core Team 2019).

3 Resultados e Discussão

Na Tabela 2 são apresentados os algoritmos de aprendizado de máquina e os respectivos valores Kappa obtidos no processo de validação cruzada, baseando-se nas 30 repetições realizadas. A partir dos valores de Kappa foi elaborado o ranking dos algoritmos, de modo que em todas as repetições os algoritmos apresentaram o mesmo comportamento.

Tabela 1 Algoritmos de aprendizado de máquina e os respectivos parâmetros utilizados no processamento da imagem.

Algoritmo	Sigla	Tipo de parâmetro	Parâmetro
<i>K-Nearest Neighbor</i>	kNN	k: identificados por uma medida de distância, a qual compara os vetores de características marcada e o conjunto de instâncias de treinamento obtidas pelo classificador	1
<i>Random Forest</i>	RF	<i>mtry</i> : número de variáveis amostradas aleatoriamente <i>ntree</i> : número de árvores	Sqrt(p) 500
<i>Support Vector Machine</i>	SVM	<i>kernel</i> : tipo de função kernel utilizada na predição, pode ser Radial, Linear, Polinomial e Sigmoidal <i>C</i> : ajusta a sensibilidade da margem de decisão de vetores de suporte classificados errados	Radial 1
<i>Artificial Neural Networks</i>	ANN	<i>maxit</i> : número máximo de interações <i>weights</i> : peso para cada exemplo	100 1

Fonte: Adaptação de Souza et al. (2016).

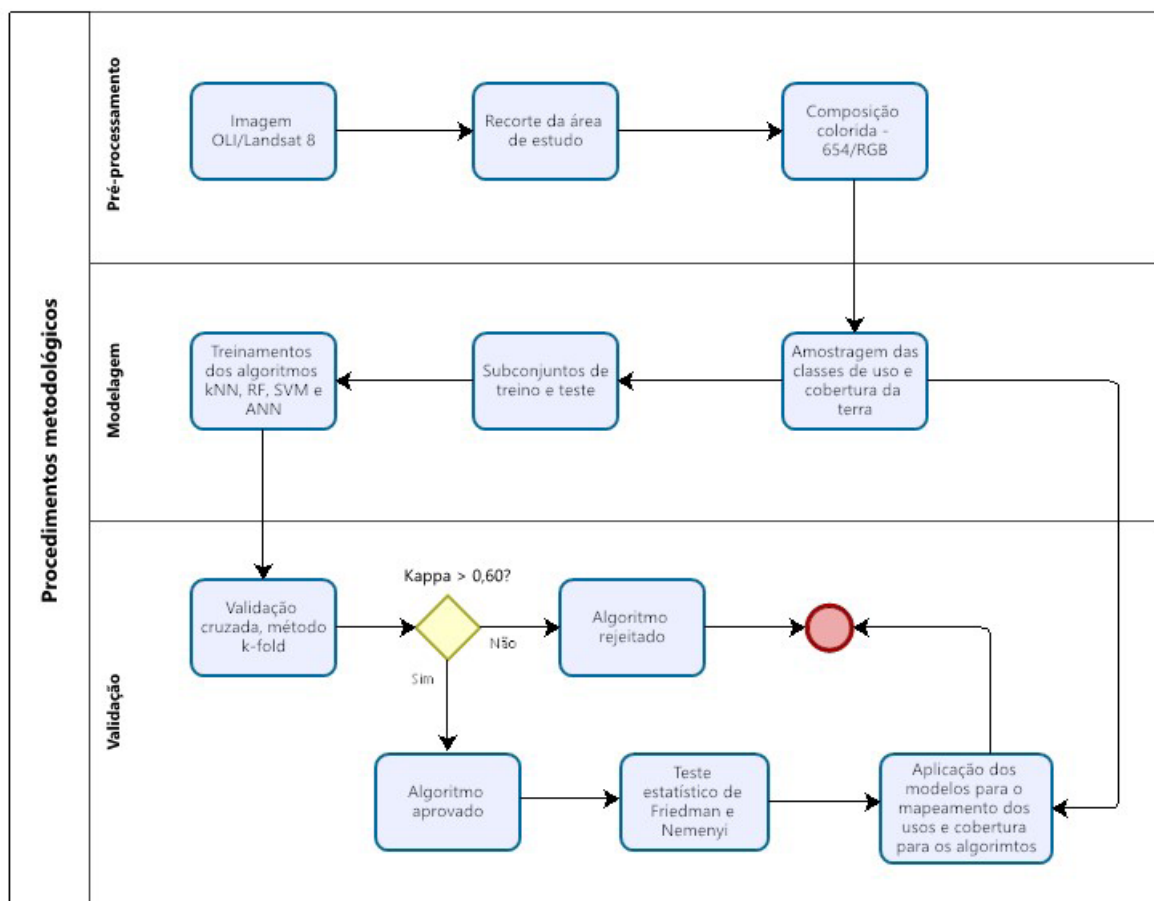


Figura 2 Fluxograma das operações de processamento desde a imagem bruta até a validação e mapeamento dos usos e cobertura da terra para os diferentes algoritmos.

Os algoritmos RF e kNN tiveram maior Kappa em todas as repetições, apresentando um coeficiente médio de 0,9841 e 0,9780, respectivamente. Na sequência, encontraram-se os algoritmos SVM e ANN, que apresentou o pior desempenho entre os algoritmos, com valores de Kappa de 0,9679 e 0,7551, respectivamente.

Corroborando esses resultados, em um estudo desenvolvido por Sothe et al. (2017), o algoritmo RF apresentou melhor desempenho com uso de imagens OLI/Landsat 8 por meio do emprego de variáveis texturais e índices de vegetação, expressando um índice Kappa de 0,88.

Tabela 2 Valores Kappa obtidos em cada repetição para os algoritmos de aprendizado de máquina RF, kNN, SVM e ANN e o respectivo ranking.

Repetições	RF	kNN	SVM	ANN	Ranking dos algoritmos			
					RF	kNN	SVM	ANN
1	0,983	0,978	0,968	0,750	1	2	3	4
2	0,985	0,980	0,969	0,764	1	2	3	4
3	0,984	0,975	0,968	0,813	1	2	3	4
4	0,985	0,976	0,968	0,792	1	2	3	4
5	0,986	0,979	0,969	0,765	1	2	3	4
6	0,984	0,979	0,968	0,723	1	2	3	4
7	0,985	0,978	0,967	0,758	1	2	3	4
8	0,985	0,978	0,968	0,776	1	2	3	4
9	0,984	0,978	0,968	0,724	1	2	3	4
10	0,982	0,978	0,968	0,740	1	2	3	4
11	0,984	0,979	0,967	0,761	1	2	3	4
12	0,985	0,977	0,967	0,747	1	2	3	4
13	0,985	0,979	0,968	0,704	1	2	3	4
14	0,985	0,978	0,968	0,714	1	2	3	4
15	0,985	0,980	0,968	0,706	1	2	3	4
16	0,984	0,978	0,968	0,770	1	2	3	4
17	0,983	0,978	0,968	0,778	1	2	3	4
18	0,983	0,977	0,969	0,673	1	2	3	4
19	0,985	0,977	0,968	0,801	1	2	3	4
20	0,984	0,979	0,968	0,771	1	2	3	4
21	0,983	0,978	0,968	0,762	1	2	3	4
22	0,983	0,979	0,968	0,748	1	2	3	4
23	0,984	0,977	0,968	0,728	1	2	3	4
24	0,983	0,977	0,967	0,810	1	2	3	4
25	0,985	0,977	0,967	0,816	1	2	3	4
26	0,985	0,979	0,968	0,748	1	2	3	4
27	0,982	0,978	0,967	0,727	1	2	3	4
28	0,985	0,978	0,968	0,739	1	2	3	4
29	0,983	0,977	0,968	0,760	1	2	3	4
30	0,985	0,979	0,967	0,784	1	2	3	4
Valor médio	0,9841	0,9780	0,9679	0,7551	1,00	2,00	3,00	4,00

Em outro estudo, realizado por Oliveira (2019), o algoritmo kNN obteve os melhores resultados, em que cinco das oito vezes apresentou maior acurácia, o algoritmo SVM ficou em segundo lugar, melhor em quatro oportunidades, e o RF obteve resultado superior em apenas uma oportunidade.

Foi aplicado o teste estatístico de Friedman e Nemenyi para verificar existência de diferenças significativas entre os algoritmos por meio da observação dos valores de Kappa (Figura 3). A distância crítica do teste foi de 0,856, de modo que todos os algoritmos apresentam uma distância superior a esse número, diferindo uns dos outros a um

nível de significância de 5%. Dessa forma, recomenda-se o algoritmo RF para a identificação dos usos presentes na Caatinga, por apresentar o melhor desempenho quando comparado aos demais algoritmos.

A Figura 4 demonstra a espacialização do uso e cobertura da terra para os diferentes algoritmos de aprendizado de máquina, nos quais os resultados demonstram a similaridade nos dados obtidos, com predominância da vegetação florestal nativa. As áreas agrícolas estão situadas, em sua maior parte, ao Sul de Petrolina, juntamente com as áreas de solo exposto, próximas ao curso do Rio São Francisco no qual serve de base para a agricultura irrigada.

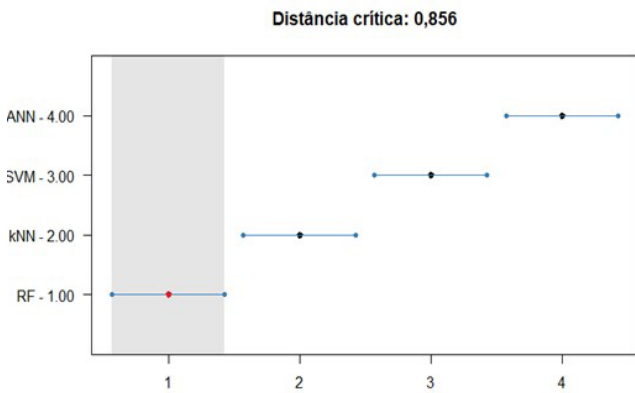


Figura 3 Teste estatístico de Friedman e Nemenyi para os algoritmos de aprendizado de máquina RF, kNN, SVM e ANN ao nível de significância de 5%.

O solo exposto corresponde, em sua maioria, a solos agrícolas em preparo para a implantação da nova cultura, como detectado por Silva et al. (2019). Os autores avaliaram alguns parâmetros biofísicos, como Albedo e NDVI, e verificaram que as regiões com pomares recém-plantados apresentavam um comportamento semelhante ao solo exposto em virtude da maior exposição do solo. Além disso, o solo exposto também pode ser atribuído ao modelo de sequeiro que é praticada na região, no qual cultiva-se

no período chuvoso (de janeiro a abril) e o restante do ano o solo fica exposto (Castro & Santos 2015).

A quantificação do mapeamento obtido pelos algoritmos de aprendizado de máquina pode ser visualizada na Tabela 3. Destaca-se que a organização dos dados se deu a partir do ranking do melhor algoritmo. Observou-se a similaridade nos valores de área para a classe de vegetação florestal nos algoritmos RF, KNN e SVM. Os algoritmos kNN e SVM apresentaram maior confusão na classe nuvem/sombra, o que contribuiu para a redução dos valores de Kappa e consequentemente, perda da qualidade das classificações.

O algoritmo ANN mostrou-se inferior aos demais, uma vez que não foi possível a identificação da classe nuvem/sombra, sendo que essas áreas foram classificadas como Água, gerando um aumento significativo dessa classe. Do mesmo modo, em um estudo desenvolvido por Chagas, Carvalho e Bhering (2011) houve discordância entre o mapa convencional de solos e o mapa de solos digital em virtude da deficiência no aprendizado dos critérios utilizados no treinamento da ANN, deficiências das variáveis ambientais utilizadas em representar todas as variações das características ambientais das unidades de mapeamento. França (2007) também observou confusão na classificação das classes de pastagem conservada, pastagem degradada e área urbana, demonstrando que o algoritmo ANN se mostra pouco eficiente nas classificações.

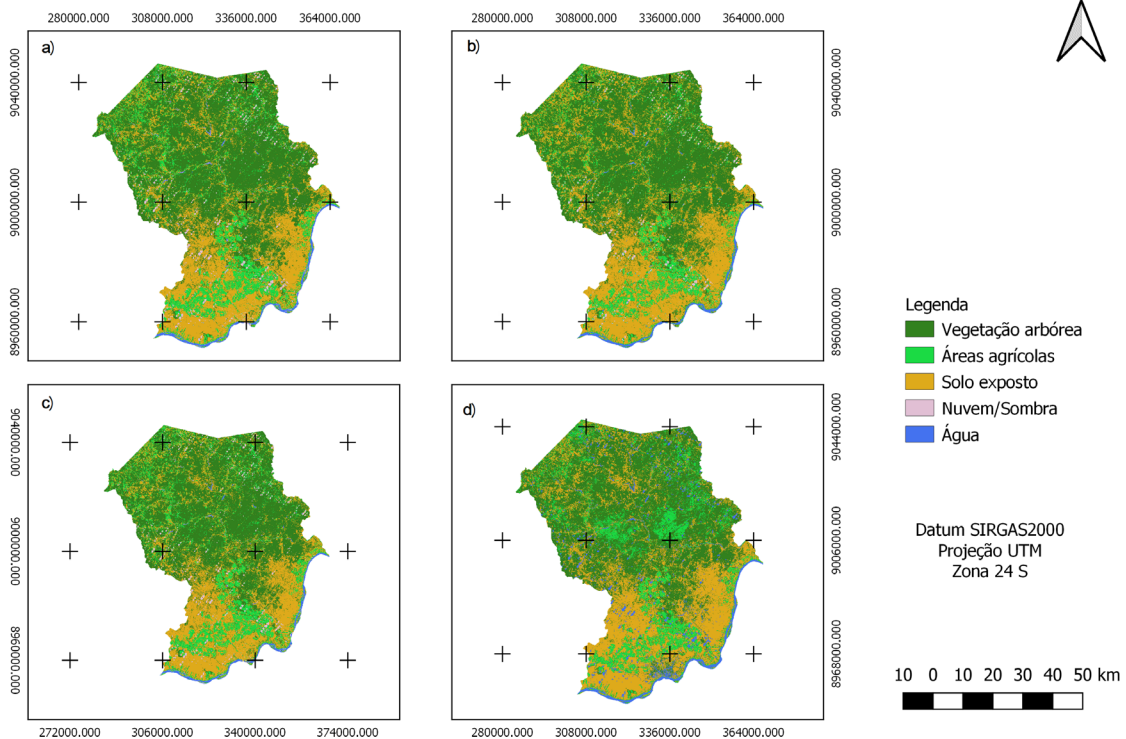


Figura 4 Mapeamento do uso e cobertura da terra a partir dos algoritmos: A. SVM; B. kNN; C. RF; D. ANN.

Na Tabela 4 encontram-se os valores da matriz de confusão obtida para o algoritmo de melhor desempenho nos valores de Kappa, o RF. Assim, as colunas referem-se à classe estimada no processo de classificação e as linhas a verdade terrestre, em que a diagonal principal representa previsões corretas e a demais indicam as incorretas.

A partir da análise da matriz de confusão, observou-se que a identificação dos usos por meio do RF englobou maiores erros na classificação da vegetação florestal, quando comparada as demais classes de uso e cobertura da terra presentes em áreas de Caatinga. Ainda que 96,11%

dos pixels foram classificados corretamente, a vegetação florestal apresentou confusão com solo exposto (3,33%) e áreas agrícolas (0,56%). A confusão gerada na classificação dessas classes pode ser justificada pela similaridade na resposta espectral, especialmente nas áreas onde a vegetação é arbustiva, uma vez que ocorre maior influência do solo na intensidade do sinal armazenado pelo sensor OLI. Esse comportamento é corroborado ao observar a classe solo exposto, a qual apresentou um percentual de 3,26% dos pixels como sendo classificados de forma errônea como vegetação florestal.

Tabela 3 Quantificação das classes de uso e cobertura da terra, em hectares, obtidos pelos algoritmos de aprendizado de máquina RF, kNN, SVM e ANN.

Classes de uso	RF	kNN	SVM	ANN
Vegetação florestal	237.135,27	237.111,69	239.422,15	217.745,94
Áreas Agrícolas	53.855,00	66.293,80	63.038,96	64.761,09
Solo exposto	150.492,18	135.168,73	134.091,43	150.029,93
Nuvem/Sombra	7.219,04	11.277,72	13.499,45	Não identificado
Água	7.458,49	6.308,04	6.108,00	23.623,03
Total	456.159,99	456.159,99	456.159,99	456.159,99

Tabela 4 Matriz de confusão para a classificação supervisionada para o algoritmo RF.

	Vegetação florestal	Áreas Agrícolas	Solo exposto	Nuvem/Sombra	Água
Vegetação florestal	346 (96,11%)	0 (0,00%)	10 (3,26%)	0 (0,00%)	0 (0,00%)
Áreas Agrícolas	2 (0,56%)	340 (100%)	0 (0,00%)	0 (0,00%)	0 (0,00%)
Solo exposto	12 (3,33%)	0 (0,00%)	300 (96,77%)	0 (0,00%)	0 (0,00%)
Nuvem/Sombra	0 (0,00%)	0 (0,00%)	0 (0,00%)	288 (100%)	0 (0,00%)
Água	0 (0,00%)	0 (0,00%)	0 (0,00%)	0 (0,00%)	91 (100%)
TOTAL	360	340	310	288	91

Entretanto, a confusão nas classes foi irrisória, demonstrando grande ajuste do algoritmo RF na identificação, especialmente da vegetação florestal do bioma Caatinga. Destaca-se que classes de uso como as áreas agrícolas e a água, não apresentaram confusão, demonstrando ser facilmente mapeadas pela técnica de inteligência artificial e dados de média resolução espacial.

4 Conclusão

O RF corresponde ao algoritmo mais eficiente para a identificação de fisionomias em região do semiárido brasileiro, o qual apresentou melhores resultados e eficácia em comparação os demais. Por outro lado, o algoritmo ANN demonstrou pouca eficiência uma vez que não foi

capaz de identificar corretamente todas as classes temáticas, confundindo a classe água com as nuvens e sombras.

A utilização da técnica de inteligência artificial com o uso de aprendizagem de máquina traz benefícios para o mapeamento do uso e cobertura da terra, pois auxilia no acompanhamento, na quantificação e na qualificação dos elementos que recobrem a superfície terrestre. Uma vez identificada com precisão as áreas da cobertura florestal, a técnica de aprendizado de máquina (machine learning) se apresenta promissora para o desenvolvimento de pesquisas voltadas ao entendimento do comportamento e interação planta-atmosfera, contribuindo para estudos sobre as contribuições da Caatinga nos processos de manutenção da qualidade climática.

5 Referências

- Alba, E. 2020, 'Influência e análise da cobertura florestal na modificação do albedo com o uso de inteligência artificial e sensoriamento remoto', PhD thesis, Universidade Federal de Santa Maria, Santa Maria.
- Alcantara, F.C., Ferreira, L.F. & Santos, H.S. 2018, 'Avaliação do kNN para reconhecimento de placas veiculares modelo Brasileiro', *Perspectiv@*, vol. 15, pp. 6-9.
- Araújo, B.A., Neto Dantas, J., Alves, A.S. & Araújo, P.A.A. 2012, 'Estrutura fitossociológica em uma área de Caatinga no Seridó Paraibano', *Revista Educação Agrícola Superior*, vol. 27, no. 1, pp. 25-9, DOI:10.12722/0101-756X.v27n01a04.
- Castro, F.C. & Santos, A.M. 2015, 'Susceptibilidade ambiental a salinização das terras em municípios da microrregião de Petrolina-Pernambuco-Brasil', *Caminhos de Geografia*, vol. 16, no. 56, pp. 160-72.
- Chagas, C.S., Carvalho, W.Jr. & Bhering, S.B. 2011, 'Integração de dados do Quickbird e atributos do terreno no mapeamento digital de solos por redes neurais artificiais', *Revista Brasileira de Ciência do Solo*, vol. 35, no. 3, pp. 693-704, DOI:10.1590/S0100-06832011000300004.
- Demsar, J. 2006, 'Statistical comparisons of classifiers over multiple data sets', *Journal of Machine Learning Research*, vol. 7, pp. 1-30.
- Fernandes, A.M.R. 2003, *Inteligência artificial: Noções gerais*, Visual Books, Florianópolis.
- França, M.M. 2007, 'Avaliação de classificações supervisionadas com Redes Neurais Artificiais e MAXVER para caracterização do uso da terra no município de Viçosa-MG', Master thesis, Universidade Federal de Viçosa, Viçosa.
- Friedman, M. 1940, 'A comparison of alternative tests of significance for the problem of m rankings', *Annals of Mathematical Statistics*, vol. 11, no. 1, pp. 86-92, DOI:10.1214/aoms/1177731944.
- Gaiad, N.P., Martins, A.P.M., Debastiani, A.B., Corte, A.P.D. & Sanquetta, C.R. 2017, 'Uso e cobertura da terra apoiados em algoritmos baseados em aprendizado de máquina: o caso de Mariana-MG', *Enciclopédia Biosfera*, vol. 14, no. 25, pp. 1211-20, DOI:10.18677/EnciBio_2017A99.
- Gariglio, M.A., Sampaio, E.V.S.B., Cestaro, L.A. & Kageyama, P.Y. 2010, *Uso sustentável e conservação dos recursos florestais da Caatinga*, Serviço Florestal Brasileiro, Brasília.
- IBGE Instituto Brasileiro de Geografia e Estatística 2004, *Portal de Mapas*, visualizado 22 Novembro 2021, <<http://mapas.ibge.gov.br/biomas2/viewer.htm>>.
- Landis, J. & Koch, G.G. 1977, 'The measurements of agreement for categorical data', *Biometrics*, vol. 33, no. 1, pp. 159-74, DOI:10.2307/2529310.
- Marchesan, J., Alba, E., Schuh, M.S., Favarin, J.A.S. & Pereira, R.S. 2020, 'Aboveground biomass estimation in a tropical forest with selective logging using random forest and lidar Data', *Floresta*, vol. 50, no. 4, pp. 1873-82, DOI:10.5380/rf.v50.i4.66589.
- Montaño, R.A.N.R., Sanquetta, C.R., Wojciechowski, J., Matta, E., Corte, A.P.D. & Todt, E. 2017, 'Artificial intelligence models to estimate biomass of tropical forest trees', *Polibits*, vol. 56, pp. 29-37.
- Moreira, M.A. 2005, *Fundamentos do sensoriamento remoto e metodologias de aplicação*, UFV, Viçosa.
- Nascimento, R.F.F., Alcântara, E.H., Kampel, M., Stech, J.L., Novo, E.M.L.M. & Fonseca, L.M.G. 2009, 'O algoritmo Support Vector Machines (SVM): avaliação da separação ótima de classes em imagens CCD-CBERS-2', *Anais XIV Simpósio Brasileiro de Sensoriamento Remoto*, INPE, Natal, 25-30 Abril, vol. 14, pp. 2079-86.
- Nemenyi, P.B. 1963, 'Distribution-free multiple comparisons', PhD thesis, Princeton University, New Jersey.
- Oliveira, P.D.S. 2019, 'Uso de aprendizagem de máquina e redes neurais convolucionais profundas para a classificação de áreas queimadas em imagens de alta resolução espacial', Master thesis, Universidade de Brasília, Brasília.
- Qgis Development Team 2019, *QGIS Geographic Information System*, Open Source Geospatial Foundation Project, visualizado 22 Dezembro 2019, <<http://www.qgis.org/>>.
- Qian, Y., Zhou, W., Yan, J., Li, W. & Han, L. 2015, 'Comparing machine learning classifiers for object-based land cover classification using very high resolution imagery', *Remote Sensing*, vol. 7, no. 1, pp. 153-68, DOI:10.3390/rs70100153.
- R Development Core Team 2019, *R: a language and environment for statistical computing*, R Foundation for Statistical Computing, Vienna, visualizado 22 Dezembro 2019, <<http://www.R-project.org/>>.
- Rodríguez, G.V. & Chica, R.M. 2012, 'Clasificación de imágenes de satélite mediante software libre: nuevas tendencias en algoritmos de Inteligencia Artificial', *XV Congreso Nacional de Tecnologías de la Información Geográfica*, AGE-CSIC, Madrid, pp. 19-21.
- Silva, D.A.O., Lopes, F.M.O., Moura, G.B.A., Silva, E.F.F., Silva, J.L.B. & Bezerra, A.C. 2019, 'Evolução espaço-temporal do risco de degradação da cobertura vegetal de Petrolina-PE', *Revista Brasileira de Meteorologia*, vol. 34, no. 1, pp. 89-99, DOI:10.1590/0102-7786334018.
- Silva, F.H.B.B., Silva, M.S.L., Cavalcanti, A.C. & Cunha, T.J.F. 2006, 'Principais solos do Semiárido do Nordeste do Brasil: "Dia de Campo"', in *Embrapa Semiárido-Artigo em anais de congresso*, Embrapa, Juazeiro, pp. 2-35.
- Soares, D.G. & Teive, R.C.G. 2015, 'Previsão de cheias do Rio Itajaí-Açu utilizando redes neurais artificiais', *Anais do Computer on the Beach*, pp. 308-17, DOI:10.13140/RG.2.1.4414.9366.
- Souza, C.G., Carvalho, L., Aguiar, P. & Arantes, T.B. 2016, 'Algoritmos de aprendizagem de máquina e variáveis de sensoriamento remoto para o mapeamento da cafeicultura', *Boletim de Ciências Geodésicas*, vol. 22, no. 4, pp. 751-73, DOI:10.1590/S1982-21702016000400043.
- SONDA Sistema de Organização Nacional de Dados Ambientais 2021, *Rede Sonda*, visualizado 25 Novembro 2021, <http://sonda.ccst.inpe.br/estacoes/petrolina_clima.html>.
- Sothe, C., Liesenberg, V., Almeida, C.M. & Schimalski, M.B. 2017, 'Abordagens para classificação do estágio sucessional da vegetação do parque nacional de São Joaquim empregando imagens landsat-8 e rapideye', *Boletim de Ciências Geodésicas*, vol. 23, no. 3, pp. 389-404, DOI:10.1590/S1982-21702017000300026.

Taura, T.A., Alvarez, I.A., Sá, I.B., Pereira, L.A. & Santos, S.M. 2011, 'Sensoriamento remoto na análise da expansão do uso e ocupação do solo em Petrolina-PE', *Anais XV Simpósio Brasileiro de Sensoriamento Remoto*, INPE, Curitiba, 30 de abril a 5 de maio, pp. 6939-46.

Zheng, S., Cao, C., Dang, Y., Xiang, H., Zhao, J., Zhang, Y., Wang, X. & Guo, H. 2013, 'Retrieval of forest growing stock volume by two different methods using Landsat TM images', *International Journal of Remote Sensing*, vol. 35, no. 1, pp. 29-43, DOI:10.1080/01431161.2013.860567.

Contribuições dos Autores

Elisiane Alba: Conceituação; metodologia; redação – revisão e edição. **Marta Laura de Souza Alexandre:** Análise formal; validação; redação - rascunho original; visualização. **Juliana Marchesan:** Conceituação; metodologia. **Luciana Sandra Bastos de Souza:** Metodologia; supervisão. **Alan César Bezerra:** Redação - revisão e edição; análise formal. **Emanuel Araújo Silva:** Conceituação; supervisão.

Conflito de interesse

Os autores declaram nenhum conflito de interesse.

Declaração de disponibilidade de dados

Scripts e códigos estão disponíveis mediante solicitação.

Como citar:

Alba, E., Alexandre, M.L.S., Marchesan, J., Souza, L.S.B., Bezerra, A.C. & Silva, E.A. 2022, 'Comparação entre Algoritmos de Aprendizado de Máquina para a Identificação de Floresta Tropical Sazonalmente Seca', *Anuário do Instituto de Geociências*, 45:40758. https://doi.org/10.11137/1982-3908_45_40758

Financiamento

Apoio da Fundação de Amparo à Ciência e Tecnologia do Estado de Pernambuco – FACEPE, projeto nº BIC-0710-5.02/21 por meio de bolsa de iniciação científica.

Editora chefe

Dra. Claudine Dereczynski

Editor Associado

Dr. Marcus Vinícius Alves de Carvalho