UNIVERSIDADE FEDERAL DO RECÔNCAVO DA BAHIA
CENTRO DE CIÊNCIAS AGRÁRIAS AMBIENTAIS E BIOLÓGICAS
PROGRAMA DE PÓS GRADUAÇÃO EM SOLOS E QUALIDADE DE ECOSSISTEMAS

DESEMPENHO DE ALGORITMOS DE APRENDIZADO DE MÁQUINA PARA MAPEAMENTO DIGITAL DE SOLOS EM ÁREA DE TABULEIROS INTERIORANOS

FELIPE TORRES SAMPAIO

CRUZ DAS ALMAS - BAHIA
MAIO – 2022

DESEMPENHO DE ALGORITMOS DE APRENDIZADO DE MÁQUINA PARA MAPEAMENTO DIGITAL DE SOLOS EM ÁREA DE TABULEIROS INTERIORANOS

FELIPE TORRES SAMPAIO

Engenheiro Florestal

Universidade Federal do Recôncavo da Bahia, 2019

Dissertação apresentada ao colegiado do Programa de Pós-Graduação em Solos e Qualidade de Ecossistemas da Universidade Federal do Recôncavo da Bahia como requisito parcial para obtenção do título de Mestre em Ciência do Solo.

Orientador: Dr. Francisco Alisson Da Silva Xavier

CRUZ DAS ALMAS - BAHIA
MAIO – 2022

FICHA CATALOGRÁFICA

S192d

Sampaio, Felipe Torres.

Desempenho de algoritmos de aprendizado de máquina para mapeamento digital de solos em área de tabuleiros interioranos / Felipe Torres Sampaio._ Cruz das Almas, BA, 2022.

50f.; il.

Dissertação (Mestrado) – Universidade Federal do Recôncavo da Bahia, Centro de Ciências Agrárias, Ambientais e Biológicas, Mestrado em Solos e Qualidade de Ecossistemas.

Orientador: Dr. Francisco Alisson Da Silva Xavier

1.Solo – Manejo. 2.Levantamento no solo – Inovações tecnológicas – Análise. I.Universidade Federal do Recôncavo da Bahia, Centro de Ciências Agrárias, Ambientais e Biológicas. II.Título.

CDD: 631.4

Ficha elaborada pela Biblioteca Universitária de Cruz das Almas - UFRB. Responsável pela Elaboração – Antonio Marcos Sarmento das Chagas (Bibliotecário - CRB5 / 1615).

UNIVERSIDADE FEDERAL DO RECÔNCAVO DA BAHIA
CENTRO DE CIÊNCIAS AGRÁRIAS AMBIENTAIS E BIOLÓGICAS
PROGRAMA DE PÓS GRADUAÇÃO EM SOLOS E QUALIDADE DE ECOSSISTEMAS

DESEMPENHO DE ALGORITMOS DE APRENDIZADO DE MÁQUINA PARA MAPEAMENTO DIGITAL DE SOLOS EM ÁREA DE TABULEIROS INTERIORANOS

COMISSÃO EXAMINADORA DA DEFESA DE DISSERTAÇÃO DE FELIPE TORRES SAMPAIO

gov.br Documento assinado digitalmente
FRANCISCO ALISSON DA SILVA XAVIER
Data: 08/02/2023 14:47:46-0300
Verifique em https://verificador.iti.br

Prof. Dr. Francisco Alisson Da Silva Xavier
Embrapa Mandioca e Fruticultura

Prof. Dr. Elpídio Inácio Fernandes Filho
Universidade Federal de Viçosa

gov.br Documento assinado digitalmente
EVERTON LUIS POELKING
Data: 09/02/2023 10:51:32-0300
Verifique em https://verificador.iti.br

Prof. Dr. Everton Luís Poelking
Universidade Federal do Recôncavo da Bahia

Dissertação homologada pelo Colegiado do Curso de Mestrado em Solos e Qualidade de Ecossistemas em _____ , conferindo o Grau de Mestre em Solos e Qualidade de Ecossistemas em_____ .

DEDICO

A meus pais, Ilná Nascimento Torres e Elmo Antônio Ribeiro Sampaio, minha irmã Amanda Torres Sampaio e Deus.

“O sol está brilhando e o clima é agradável...”

(Bob Marley)

AGRADECIMENTOS

Gostaria de agradecer a oportunidade de realizar uma pós-graduação em uma universidade pública, pois acredito que a educação é a chave para o desenvolvimento da humanidade. Durante o período de desenvolvimento do trabalho o mundo se encontra em um momento muito delicado, onde, por conta da pandemia originada pela ocorrência do Covid-19, muitos valores foram ressignificados. Assim, dedico também esse trabalho a todos que participaram de forma positiva na luta contra a pandemia.

Agradecer a oportunidade de estar vivo, respirar, sentir o sol, ter a cognição, conseguir raciocinar, tudo isso é motivo para agradecer, afinal de contas, o Sol está brilhando, permitindo um novo dia chegar. Cada dia é uma oportunidade que recebemos de melhorar e evoluir. Após tantas dificuldades, agradeço a vida e pela vida de todas as pessoas que amo.

Gostaria de agradecer a equipe principal que realizou esse levantamento, ao professor Oldair e ao professor Luciano, por todas as aulas e ensinamentos que me foram passados, aos conhecimentos de campo, aos trabalhos desenvolvidos de laboratório, pela disponibilidade e paciência e, principalmente, como exemplo de profissionais e humanos que são; agradeço ao engenheiro Icaro pelas trilhas de levantamento de campo, estudos de novos softwares e tecnologias, e parceria nos esportes, na música e no laboratório. Quero agradecer ao professor Everton e a professora Isabel pelas dicas e contribuições durante o período qualificação.

Registro também meu agradecimento a todos os professores das disciplinas cursadas que foram fundamentais para o desenvolvimento do projeto, em especial o professor Elpídio por todas as contribuições, ensinamentos e disponibilidade para ajudar no meu desenvolvimento.

A todos os funcionários que fazem a universidade funcionar, que ajudam o serviço público brasileiro, deixo meus mais sinceros agradecimentos.

Agradeço aos meus amigos, que sempre me incentivaram, muito obrigado.

SUMÁRIO

1. LISTA DE FIGURAS

2. LISTAS DE TABELAS E QUADROS

TABELAS

QUADROS

Desempenho de algoritmos de aprendizado de máquina para mapeamento digital de solos em área de Tabuleiros Interioranos

RESUMO

A gestão do solo requer conhecimento sobre as propriedades e atributos que os caracterizam e o mapeamento deste recurso se constitui em importante ferramenta para o planejamento do seu uso e ocupação. Atualmente, utiliza-se o aprendizado de máquina (ML) no mapeamento digital de solos (MDS) para aprimorar e auxiliar o processo de obtenção de informações. Diante do exposto, o objetivo deste trabalho foi utilizar repetição de processos de classificação, variando os conjuntos teste e treino, para investigar o desempenho de três algoritmos de aprendizado de máquina na predição de classes de solo. A área selecionada está localizada na região de Tabuleiros Interioranos do Recôncavo da Bahia, Brasil. Os dados utilizados para treinar os modelos foram obtidos de 38 perfis de solos descritos até o quarto nível taxinômico e mais 150 pontos de observação que foram amostrados, utilizando o hipercubo latino condicionado. As variáveis foram geradas por meio de mapas de material origem; uso do solo; dados geomorfométricos derivados do modelo digital de elevação e imagens de satélite. No software R, modelos foram treinados utilizando validação cruzada. Os algoritmos testados foram: *Support Vector Machine Kenel Linear* (SVM), *Random Forest* (RF) e *Decision Trees C5.0* (C5). Em geral, todos os classificadores testados apresentaram grande potencial para predição de classes de solos em áreas de tabuleiros interioranos. Observou-se que a mediana da acurácia foi de 0,68 e 0,62 para os modelos C5 e RF, com desvio padrão de 0,07 e 0,08, respectivamente; enquanto que o SVM apresentou uma mediana de 0,51, com desvio padrão de 0,08. Utilizar repetições para avaliar o desempenho de algoritmos de aprendizado de máquina no mapeamento digital de solos é fundamental, uma vez que uma única repetição pode superestimar ou subestimar o verdadeiro potencial do classificador. As covariáveis mais importantes para a predição foram *Standardized height* (altura padronizada) e *effective air flow heights* (altura efetiva do fluxo de ar). A distribuição das classes de solo de solo do campus da universidade é marcada pela predominância dos Latossolos Amarelos nos topos planos dos tabuleiros e à medida que o terreno começa passa a ser ondulado outras classes se fazem presentes como é o caso do Cambissolo, Argissolo, Planossolo e Chernossolo, próximo à linha de drenagem encontramos o Vertissolo e o Gleissolo.

Palavras chaves: predição de classes de solo; teste de algoritmo; pedometria; planejamento do uso da terra.

Performance of machine learning algorithms for digital soil mapping in Inland Tablelands area

ABSTRACT

Soil management requires knowledge about the properties and attributes that characterize them, and the mapping of this constitutes an important tool for planning its use and occupation. Currently, machine learning (ML) is used in digital soil mapping (MDS) to improve and assist the information storage process. The objective of this work was repeated classification processes, varying the test and training sets, to investigate the performance of three machine learning animals in the prediction of soil classes. The selected area is located in the region of Tabuleiros Interioranos of the Recôncavo da Bahia, Brazil. The data used for the suitable models were 38 obtained soil profiles described to the taxonomic level and more than 150 observation points that were sampled, using the conditioned Latin hypercube. The consequences were caused through material origin maps; use of the soil; geomorphometric data, digital image model results and satellite imagery. In the R software, models were trained using cross-validation. The organisms tested were: Support Vector Machine Linear Kernel (SVM), Random Forest (RF) and Decision Trees C5.0 (C5). In general, all tested envoys have great potential for predicting soil classes in inland plateau areas. It is observed that the median accuracy was 0.68 and 0.62 for the C5 and RF models, with standard deviation of 0.07 and 0.08, respectively; while the SVM presented a median of 0.56, with a standard deviation of 0.08. Resource for use as a single machine can overestimate or underestimate the true potential of the fundamental ad. The most important covariates for the prediction were Standardized Height and Effective Airflow Height. The distribution of soil classes on the university campus is marked by the predominance of Latossolos Amarelos on the flat tops of the boards and as the terrain begins to become undulating, other classes are present, such as Cambissolo, Argissolo, Planossolo and Chernossolo, next to the solution line found the Vertissolo and Gleissolo.

Keywords: soil class prediction; algorithm test; pedometrics; land use planning.

1. INTRODUÇÃO

Os solos são formados por uma interação de fatores sendo eles, clima, relevo, material de origem, organismos e tempo (JENNY, 1941). Todos esses fatores influenciam nas suas características químicas, físicas, mineralógicas e consequentemente na sua classificação. É crescente o reconhecimento da importância do solo e dos serviços que ele fornece. Os solos influenciam na tomada de decisão no gerenciamento de recursos ambientais, uso e ocupação da terra, além da manutenção da regulação do clima global (TEN CATEN et al., 2012), por isso o mapeamento de sua distribuição espacial é de grande importância.

No mapeamento de solos convencional o pedólogo em campo relaciona fatores ambientais como clima, relevo, vegetação, litologia além de experiências anteriores e análises laboratoriais de amostras para determinar as classes do solo. A amostragem é a etapa inicial no processo de mapeamento de solo, e esta é realizada de forma aleatória em pontos no relevo, determinadas pela experiência do pedólogo (KEMPEN et al., 2012), mas essa amostragem pode não representar todas as variações presentes no ambiente.

O MDS é uma técnica que por meio de modelos matemáticos integram covariáveis ambientais que influenciam nos fatores de formação do solo para realizar a classificação (BAGATINI; GIASSON; TESKE, 2015).

Uma das principais fontes de informação para o MDS é o Modelo Digital de Superfície (MDE), pois o relevo é um dos principais fatores de formação do solo e a partir dele é possível determinar variáveis ambientais que influenciam no processo de formação do solo (TEN CATEN et al., 2012).

Na região dos Tabuleiros Costeiros e Interioranos, distribuídos ao longo ou próximo ao litoral nordeste brasileiro, a distribuição dos solos na paisagem é condicionada às variações geológicas locais e, principalmente as variações no relevo, uma vez que as características de vegetação e clima locais são muito semelhantes.

Algoritmos de ML supervisionados são amplamente utilizados em MDS, sua metodologia consiste em pontos amostrais de classes de solos conhecidas para treinar os algoritmos. Não existe ainda uma forma segura de determinar o número de amostras necessárias para a realização de uma classificação. Os

pontos amostrais são utilizados juntamente com covariáveis ambientais de mesma localização, esses dados são treinados dentro de uma área limite. Após treinamento do modelo, este pode ser utilizado em locais que não se tem a informação da classe de solo, apenas das covariáveis dentro da área limite do treino. (MINASNY; MCBRATNEY, 2006).

Vários algoritmos de ML têm sido usados/testados em geomorfologia, como é o caso da árvore de decisão C5.0 (C5), *Random Forest* (RF) e Máquinas de Vetores de Suporte (SVM) (Siqueira R., 2021). Para classificação de solos, preditores baseados em árvore de decisão são os mais populares, em 80% dos estudos de caso os pesquisadores usaram pelo menos um algoritmo da família das árvores de decisão como RF e C5. (WADOUX, et al. 2020). O método e a densidade de amostragem são muito variáveis na literatura, geralmente o nível de detalhe do mapeamento e tamanho da área são os principais fatores que influenciam na quantidade de amostras que são coletadas.

Para a obtenção de estimativas estáveis e reduzir as incertezas do desempenho do modelo, este é executado repetidas vezes criando diversas divisões dos dados de treinamento e teste (KUHN; JOHNSON, 2013).

Diante do exposto, o objetivo deste trabalho foi utilizar repetição de processos de classificação, variando os conjuntos teste e treino, para investigar o desempenho de três algoritmos de ML na predição de classes de solo na região de Tabuleiros Interioranos do Recôncavo da Bahia, Brasil.

2. MATERIAL E MÉTODOS

2.1 Características físico-ambientais da Área de estudo

A área selecionada para o estudo está inserida na região do Recôncavo da Bahia, se trata dos limites da Universidade Federal do Recôncavo da Bahia (UFRB), mais precisamente do campus de Cruz Das Almas, com área de 1.367 hectares (Figura 1).

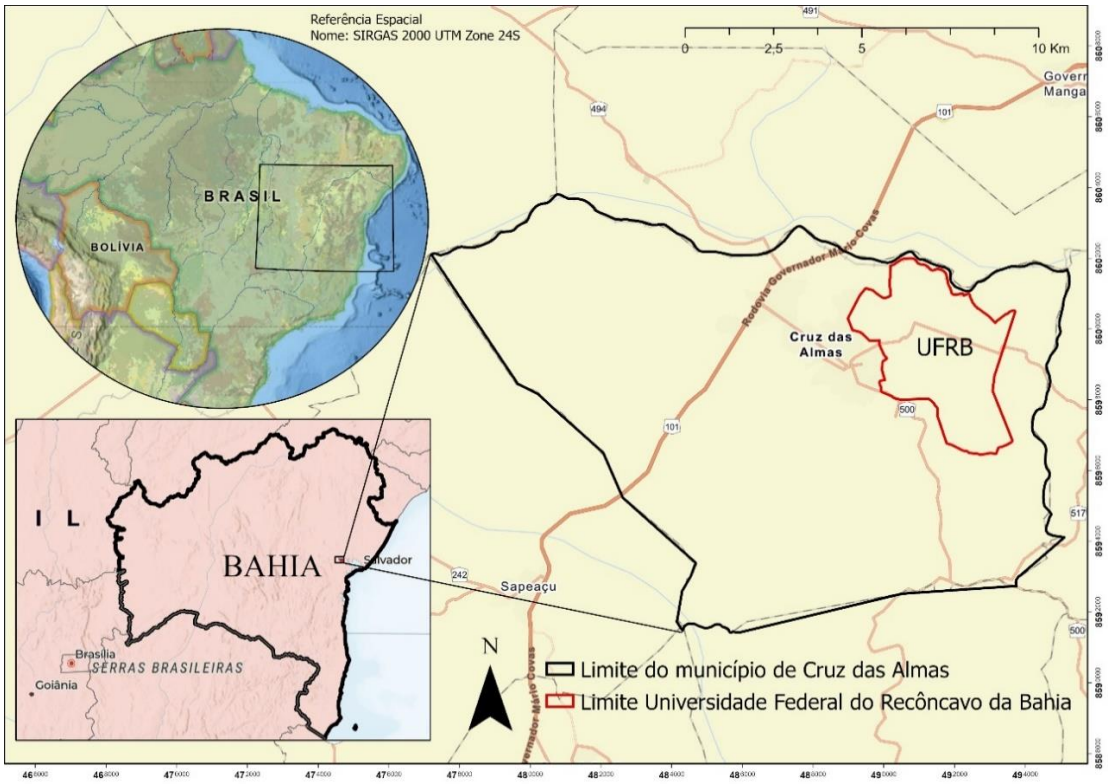

Figura 1 - Mapa de localização da área de estudo.

.

Na região ocorrem três unidades geológicas distintas: rochas metamórficas pré-cambrianas do Complexo Granulítico; depósitos detrito lateríticos terciários e sedimentos aluvionares quaternários.

O Complexo Granulítico, de idade Pré Cambriana Inferior (180 milhões de anos), apresenta rochas de alto grau de metamorfismo termo-mecânico, tais como: charnoquitos ácidos e intermediários e tipos básicos como piroxênio, granulitos além de biotita gnaisse (MASCARENHAS et al. 1979).

Os depósitos detrito lateríticos, de idade neógena (Neo-Terciário), têm espessura média de 20 a 30 metros e assenta-se sobre rochas pré-cambrianas

do Complexo Granulítico. É constituída geneticamente por sedimentos continentais: aluviais e fluviais, depositados sob clima semiárido. Litologicamente é formada por sedimentos detríticos, mal consolidados: arenitos amarelos e amarelo-avermelhados, sílticos-argilosos, mal selecionados e também siltitos argilosos. Na parte da base podem ocorrer leitos descontínuos de conglomerados de seixos de quartzo arredondados (CPRM, 2001).

Os depósitos aluvionares e/ou aluviocoluvionares e detríticos quaternários constam de intercalações de sedimentos terrígenos, inconsolidados, compostos por areias finas a médias, siltes e argilas (RIBEIRO, 1998).

Geomorfologicamente a área se apresenta principalmente tabuleiros, em fase de dissecação, classificada na unidade geomorfológica dos Tabuleiros Interioranos pelo RADAMBRASIL (1981), pertencendo ao domínio morfoestrutural dos Planaltos Inumados.

No município os tabuleiros são formados por depósitos sedimentares terciários de formação Capim Grosso em altitude média de 220 a 225 metros acima do nível do mar. As encostas são ravinadas com perfis variados (segundo os tipos de materiais geológicos sobre as quais estão modeladas). A parte superior das encostas, esculpidas sobre sedimentos terrígenos, normalmente forma um segmento retilíneo de declividade mais acentuada. As partes média e inferior das encostas são modeladas sobre rochas metamórficas de litologias variadas, que formam o Complexo Granulítico. Essas variações são responsáveis pelos tipos de perfis (côncavos ou convexos) segundo o RADAMBRASIL (1981).

O vale do rio Capivari e de seus principais tributários, que drenam a região, são formados por depósitos aluvionares recentes. Os riachos principais formam anfiteatros em suas cabeceiras. Os riachos secundários apresentam vales encaixados (RIBEIRO, et al., 1998).

Segundo Köppen (1928) o clima local é do tipo Af, quente, com o mês mais frio com temperatura superior a 18 °C e o mês mais seco com precipitação igual ou superior a 60 mm; a pluviosidade média anual é de 1.200 mm, sendo os meses de março a julho os mais chuvosos e outubro e janeiro os mais secos, com temperatura média anual de 24,2 °C.

A região de Cruz das Almas era coberta originalmente por Floresta Estacional Semidecidual (RADAMBRASIL, 1981). que foi em sua maioria devastada e substituída por pastagens e culturas agrícolas como: mandioca, cana-de-açúcar, milho, feijão, pequenos pomares de mangueiras, cajueiros, jaqueiras e laranjeiras.

2.2 Dados legados e amostragem

Tomando como base levantamentos de solos anteriores realizados na área (SOUZA & SOUZA, 2001; RIBEIRO et al, 1998; RIBEIRO et al, 1988; RODRIGUES et al, 2003); trabalhos de pesquisa desenvolvidos na área (SODRÉ et al, 2019) e novas descrições morfológicas, coleta e análises químicas e físicas de solos, 38 perfis completos foram espacializados e classificados até o quarto nível taxonômico, de acordo com o Sistema Brasileiro de Classificação de Solos - SiBCS (EMBRAPA, 2018).

A distribuição de pontos de amostragem foi realizada com auxílio do *software* R 4.1.1 (R Development Core Team, 2016) utilizando covariáveis ambientais através do método Hipercubo Latino Condicionado (cLHS) (Figura 2).O cLHS particiona as variáveis em intervalos iguais com base na sua distribuição e seleciona as amostras em cada um dos intervalos formando estratos, e as combinam (Ließ, M, 2020). A amostragem determinada pelo cLHS é otimizada, distribuída pela paisagem e não é agrupada, o que permite melhorar a amostragem para que a variabilidade dentro da área seja representada (MINASNY; MCBRATNEY, 2006).

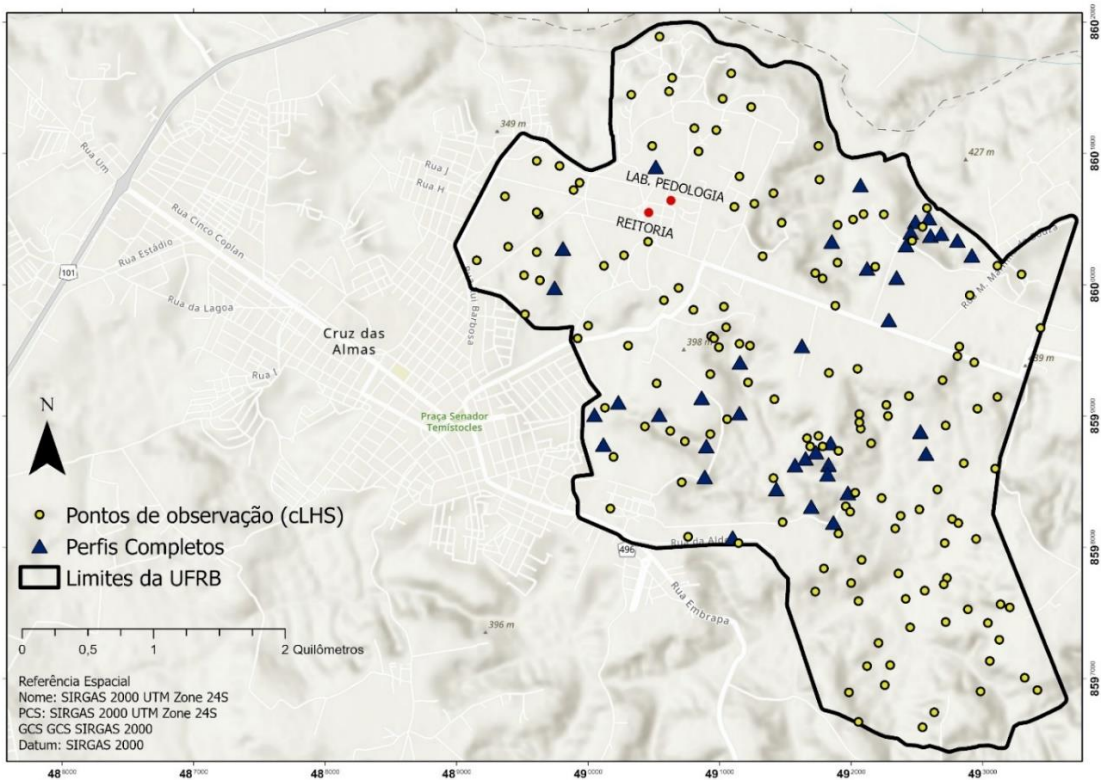

Figura 2 - Localização dos pontos de observação e perfis completos de solo na área de estudo.

Para confirmar a ocorrência dos solos anteriormente descritos e definir as unidades de mapeamento (UM), foi realizado um levantamento de campo, no período de fevereiro a agosto de 2021, onde foram identificados por um pedólogo sênior 150 pontos de observação que fazem referência aos perfis completos descritos no campus, por meio de minitrincheiras, tradagens e coleta para análise química e física. As coordenadas geográficas desses pontos de observação foram coletadas a partir de um receptor GNSS de navegação Garminn eTrex 10.

A partir deste estudo, foram identificadas 12 classes de solo e separadas em 10 unidades de mapeamento (UM) sendo elas:

- LAdx1 - LATOSSOLO AMARELO Distrocoeso típico.
- LAdx2 - LATOSSOLO AMARELO Distrocoeso argissólico
- PAdx + CXbd - Associação entre ARGISSOLO AMARELO Distrocoeso típico + CAMBISSOLO HÁPLICO Tb distrófico.
- CXve - CAMBISSOLO HÁPLICO Ta Eutrófico.

- SXe - PLANOSSOLO HÁPLICO Eutrófico.
- VGo - VERTISSOLO HÁPLICO Órtico solódico.
- MEo - CHERNOSSOLO EBÂNICO Órtico + CAMBISSOLO HÁPLICO Tb Eutrófico típicos.
- PACdx + SNo - Associação entre ARGISSOLO ACINZENTADO Distrocoeso + PLANOSSOLO NÁTRICO Órtico.
- GXve - GLEISSOLO HÁPLICO Ta Eutrófico.
- PVAe - ARGISSOLO VERMELHO AMARELO Eutrófico.

2.3 Algoritmos de aprendizado de máquina

Para predição e espacialização das unidades de mapeamento identificadas e caracterizadas anteriormente, neste trabalho, forma testados os algoritmos de classificação supervisionada *Random Forest* (RF), por meio do pacote randomForest (BREIMAN, 2006), *Support Vector Machine* Kernel Linear (SVM), utilizando o pacote kernlab (KARATZOGLOU et al., 2016) e Árvores de decisão C5.0, mediante o pacote C50 (KUHN, 2018), implementados no pacote Caret (KUHN, 2017) e executados no *software* R.

Cada algoritmo de classificação trata os dados de forma diferente, não existindo um melhor algoritmo que funcione em todas as situações (KHALEDIAN; MILLER, 2020), por essa razão foram testados 3 diferentes modelos, com 50 repetições.

O SVM de Kernel Linear (Cortes and Vapnik, 1995) separa as classes (variáveis alvo) utilizando hiperplanos em espaços n-dimensionais. Para a otimização dessa separação o algoritmo utiliza a distância máxima entre os pontos das classes e do hiperplano. No espaço bidimensional o hiperplano é uma reta que separa as duas classes (KHALEDIAN; MILLER, 2020).

RF e C5 são algoritmos desenvolvidos com base no modelo de árvore de classificação e regressão (*classification and regression tree* - CART). C5 é um dos modelos mais conhecidos de árvore de decisão. O algoritmo original é um produto comercial desenvolvido por Ross Quinlan (Quinlan, 1993). Este algoritmo usa o cálculo de entropia de informação para determinar a melhor regra que divide os dados em cada nó da árvore em classes mais puras. A cada divisão cada grupo de dados conterá diversidade de classes. O algoritmo C5 cria árvores com menores incertezas pois utiliza-se da estatística para fazer a divisão das

amostras (GUO et al., 2021), criando classes mais puras com menor diversidade a cada divisão, ou seja separa os dados (variáveis alvo) em grupos mais homogêneos/semelhantes (segundo as variáveis resposta) para realizar a classificação, além de ser um algoritmo de execução rápida e possuir bom desempenho para grandes números de dados (MURUGAN BHAGAVATHI et al., 2021).

Com o objetivo de aumentar a variabilidade das informações e de criar árvores de decisão diferentes a partir dos mesmos dados, é utilizada uma técnica de amostragem chamada *bootstrap* (amostragem com reposição). A técnica de amostrar repetidas vezes o conjunto de dados usando *bootstrap* e criar várias árvores é denominada de *bagging* (*bootstrap agreggation*). Conforme proposto por Breiman (2001), o RF é uma evolução do *bagging*. O *bagging* utiliza sempre o mesmo conjunto de variáveis para criar um conjunto de árvores e elas normalmente são correlacionadas entre si. O RF escolhe de forma também aleatória as variáveis durante a construção das árvores, amostrando um subconjunto delas. Isto gera árvores diferentes com baixa correlação e aumenta sua diversidade. Ao final para cada amostras são obtidos n resultados de classificação produzidos pelas n árvores criadas. Para cada nó ele faz uma amostragem aleatória das variáveis resposta que ele vai utilizar no processo. A classe de cada amostra é determinada pela classe mais frequente (moda da variável alvo). No caso de regressão é feita a média dos valores obtidos em cada árvore. O RF usa árvores bem desenvolvidas que geralmente tem alta variância, o uso da moda ou da média diminui essa variância na hora da predição para cada amostra.

2.4 Covariáveis ambientais

Seguindo o modelo SCORPAN, conforme descrito por (MCBRATNEY; MENDONÇA SANTOS; MINASNY, 2003), as covariáveis ambientais utilizadas nesse estudo foram derivadas de um modelo digital de superfície (MDE) gerado a partir de curvas de nível extraídas de um mapa topográfico em escala 1:25.000, adquiridas na base de dados digitais do Exército Brasileiro (BDGEX ,2019) que foram interpoladas por meio do algoritmo *Topo to raster* no ArcGIS 10.2 (ESRI, 2019) resultando em um *raster* de resolução espacial de 5 metros.

Em ambiente R com o pacote RSaga (Brenning et al., 2018) para acessar as ferramentas de *software SAGA GIS* 6.2, foram geradas 42 covariáveis geomorfométricas a partir do MDE, entre elas, declividade, sombreamento, índice de umidade topográfico, índice de radiação solar, orientação das encostas, curvatura de perfil, curvatura planar, altura padronizada, profundidade do vale e comprimento de rampa.

Imagens espectrais do satélite CBERS 04A, adquiridas em 23 de novembro de 2019: banda 1 azul (0,45 - 0,52 µm), banda 2 verde (0,52 - 0,59 µm), banda 3 vermelha (0,63 - 0,69 µm), banda 4 infravermelho (0,77 - 0,89 µm), foram obtidas por meio do Instituto Nacional de Pesquisas Espaciais (INPE), a partir do sensor WPM com resolução de 8 metros.

Os mapas de uso da terra e NDVI foram produzidos com o software *ArcGis Pro* da ESRI em resolução espacial de 8 metros. O mapa de material de origem foi adaptado do mapa de Geologia do Brasil publicado pelo Serviço Geológico do Brasil - CPRM na escala de 1: 250.000 e foi interpolado com o *software ArcGis Pro* da ESRI para uma resolução espacial de 8 metros.

Também em ambiente R foram gerados os *rasters* referentes à latitude e longitude em metros (coordenadas UTM) de cada célula. Assim, o presente trabalho contou com 53 covariáveis ambientais.

Todas as covariáveis utilizadas estão listadas na Tabela 1.

Tabela 1 - Lista de variáveis e suas resoluções originais.

Código	**Nome**	**Resolução do *Raster***
Aspect	Aspecto	5 metros
b1	Banda 1 CBRS 04A sensor WPM	8 metros
b2	Banda 2 CBRS 04A sensor WPM	8 metros
b3	Banda 3 CBRS 04A sensor WPM	8 metros
b4	Banda 4 CBRS 04A sensor WPM	8 metros
convergence_index	índice convergência	5 metros
curso	Distância da linha de drenagem	8 metros
curv_cross_sectional	Seção transversal de curvatura	5 metros
curv_flow_line	Linha do fluxo de curvatura	5 metros
curv_general	Disposição geral	5 metros
curv_longitudinal	Disposição longitudinal	5 metros
curv_maximal	Curvatura máxima	5 metros
curv_minimal	Curvatura minima	5 metros
curv_plan	Curvatura Planar	5 metros
curv_profile	Curvatura de Perfil	5 metros
curv_tangencial	Curvatura tangencial	5 metros
curv_total	Curvatura total	5 metros
curvature_classification	Classificação de curvatura	5 metros

difference	Diferença	5 metros
diurnal_anisotropic_heat	Calor anisotrópico diurno	5 metros
effective_air_flow_heights	Alturas de fluxo de ar eficazes	5 metros
gradient	Gradiente de declividade do terreno	5 metros
hill	Tamanho da encosta	5 metros
hill_idx	Índice de encosta	5 metros
landforms_tpi_based	Relevos baseados em indice de posição topográfica	5 metros
lat	Latitude em metros (Coordenadas UTM)	5 metros
litologia	Material de origem	50 metros
long	Longitude em metros (Coordenadas UTM)	5 metros
Mass_balance_index	Índice de balanço de massa	5 metros
mid_slope_position	Posição de inclinação média	5 metros
morphometric_protection_index	índice de proteção morfométrica	5 metros
MRRTF	Índice de multiresolução da planicidade do topo do cume	5 metros
MRVBF	Índice multiresolução de planicidade do fundo do vale	5 metros
ndvi	Índice de Vegetação da Diferença Normalizada	8 metros
normalized_height	Altura normalizada	5 metros
real_surface_area	Área de superfície real	5 metros
Saga_wetness_index	Índice de umidade SAGA GIS	5 metros
slope_degress	Graus de inclinação das encostas	5 metros
slope_height	Altura do declive	5 metros

slope_idx	Índice de inclinação	5 metros
standardized_height	Altura padronizada	5 metros
surface_specific_points	Pontos específicos da superfície	5 metros
terrain_ruggedness_index	Índice de robustez do terreno	5 metros
terrain_surface_classification_iwahashi	Classificação da superfície do terreno iwahashi	5 metros
terrain_surface_convexity	Convexidade da superfície do terreno	5 metros
terrain_surface_texture	Textura da superfície do terreno	5 metros
topographic_position_index	Índice de posição topográfica	5 metros
valley	Tamanho dos vales	5 metros
valley_depth	Profundidade dos vales	5 metros
valley_idx	Índice das formas dos vales	5 metros
vector_ruggedness_index	índice de robustez vetorial	5 metros
Class_terra	Classificação do uso da terra	8 metros

2.5 Eliminação de variáveis

A quantidade de variáveis explicativas que demande um menor custo computacional, é preferível, além de obedecer ao princípio da parcimônia, o processo de remoção de covariáveis têm por objetivo o desenvolvimento de um modelo de melhor compreensão (Gomes, 2019).

As 53 covariáveis foram analisadas e filtradas, primeiro as covariáveis com pouca variância (carregam pouca informação da variabilidade ambiental da área de interesse) foram identificadas com a função *nearZeroVar* através do pacote Caret e em seguida foram eliminadas do *dataset*.

Covariáveis que possuíam correlação igual ou superior a 95% com outra covariável do conjunto de dados também foram identificadas e eliminadas com a função *findCorrelation* também pertencente ao pacote Caret.

Seguindo utilizando o mesmo pacote, com a função *Recursive Feature Elimination* (RFE), foram identificadas as covariáveis que eram mais importantes para o treino de cada modelo. O algoritmo seleciona o conjunto ideal de covariáveis a partir de um modelo inicial com todas possíveis e elimina as menos importantes. O RFE é específico para cada algoritmo. As variáveis selecionadas pelo RFE foram também utilizadas na fase de predição para o mapa final.

Foi gerado um gráfico com as covariáveis que mais foram selecionadas pelo RFE para explicar a distribuição espacial das classes de solo.

2.6 Treino dos classificadores

Ao final dos processos anteriores foi formado um conjunto de dados (*dataset*) com 150 observações. Utilizando a função *createDataPartition* do pacote Caret, o *dataset* foi subdividido de forma aleatória em dois conjuntos, 75% das observações para treino e 25% para teste.

Seguindo a metodologia apresentada por Meier et al. (2018), o treinamento foi feito utilizando a função *train* do pacote Caret com validação cruzada de 10 partições e 3 repetições para calibração dos hiperparâmetros de cada algoritmo testado. Com os modelos criados foi realizada a predição do conjunto de teste (amostras que não foram utilizadas no treinamento do modelo), os resultados dessa predição foram usados para determinar os parâmetros de performance dos algoritmos.

2.7 Teste e incerteza do modelo

Os parâmetros de performance utilizados foram os índices Kappa (Landis e Koch, 1977) e a acurácia geral (Equação 1). Como é característico, os dados categóricos e a sua previsão dentro dos modelos se apresentam como dados discretos, uma classe é prevista ou não é. Como variáveis binárias, a previsão é sim ou não. A matriz de confusão é utilizada para fazer os cálculos estatísticos referente as validações dos modelos. Utilizando o conjunto de dados de teste tem-se o controle do número de acertos e erros na predição de cada classe de solo.

$$Po = \sum_{i=1}^{m} \frac{n_{ii}}{n} \quad (1)$$

Com a soma das colunas da matriz de confusão obtêm-se o número total de observações (informação do campo) para cada classe de solo. Da mesma forma, se forem somadas cada uma das linhas da matriz, obtêm-se o número total de previsões de cada classe de solo, em um ponto de classe conhecida. As previsões poderiam ter sido feitas através de qualquer tipo de modelo ou processo de classificação. Portanto, os números da matriz fornecem uma taxa de concordância da classificação. A acurácia geral (Po) é, portanto, calculada dividindo o total de predições corretas pelo total número de observações.

O coeficiente Kappa (Equação 2) é uma métrica de precisão derivada de funções da matriz de confusão, e definido como uma medida da concordância real menos a concordância por chance (Equação 3) (Congalton and Green, 2009):

$$Kappa = \frac{P_o - P_e}{1 - P_e} \quad (2)$$

Onde Po é acurácia geral e Pe (Equação 3) é a probabilidade hipotética de acordo ao acaso.

$$P_e = \sum_{i=1}^{n} \left(\frac{Somatorio\ das\ classes\ encontradas\ no\ campo_i}{Número\ total\ de\ observações}\right) x \left(\frac{Somatorio\ das\ previsões\ feitas\ pelos\ algoritmos_i}{Número\ total\ de\ observações}\right)$$

(3) Fonte: Malone, Brendan P. et al, 2017

Foram gerados tabelas e gráficos com os resultados dos índices Kappa e Acurácia.

2.8 Análises estatísticas

A seleção de variáveis explicativas pelo RFE, o treinamento e a predição de classes foram repetidos 50 vezes variando os dados de treino e teste de forma aleatória, esse não é um número definido, não existe um consenso de quantidade de repetições de processos em MDS na literatura, porém um grande volume de repetições permite maior estabilidade e confiança dos resultados, possibilitando uma análise estatística (média, desvio padrão e variância), evitando uma superestimação ou subestimação dos algoritmos. Como para cada repetição ocorre variação nos dados de entrada, os resultados preliminares são de 50 mapas com ligeira diferença de classificação cada algoritmo. A classificação é feita pixel a pixel possibilitando derivar mapas estatísticos:

- O mapa da moda: referente a classificação que ocorreu mais vezes em cada pixel durante as repetições; resultando, portanto, em um mapa mais confiável da distribuição espacial das unidades de mapeamento;
- O mapa de mudança: que traz a informação dos pixels que foram classificadas somente por uma única unidade de mapeamento, refletindo as áreas que os modelos encontraram um padrão de distribuição de uma determinada unidade de mapeamento em função das covariáveis utilizadas;
- O mapa de variância: em complemento ao mapa de mudança mostra a quantidade de unidades de mapeamento que foram atribuídas em cada pixel, mostrando as áreas que os algoritmos tiveram mais dificuldade em encontrar padrões nas variáveis explicativas.
- O mapa da frequência da moda: Traz a informação de quantas vezes a unidade de mapeamento modal foi atribuída ao pixel, informando áreas com maior incerteza na predição da classe de solo.

3. RESULTADOS E DISCUSSÃO

3.1 Filtro de variáveis

Utilizando as funções *nearZeroVar* e *findCorrelation* um total de 19 covariáveis foram eliminadas (Quadro 1) com o objetivo de seguir o princípio da parcimônia e desenvolver modelos mais leves, com capacidade de explicar a distribuição das classes de solo na paisagem.

Quadro 1 - Covariáveis Eliminadas no pré-processamento

Pouca variância	
curvature_classification	Classificação de curvatura
hill	Encosta
hill_idx	índice de encosta
morphometric_protection_index	índice de proteção da morfometria
slope_idx	Índice de inclinação
terrain_surface_texture	Textura da superfície do terreno
valley	Vale
valley_idx	índice de vale
Alta correlação	
b2	Banda 2 CBRS 04A sensor WPM
b3	Banda 3 CBRS 04A sensor WPM
convergence_index	índice convergência
curv_cross_sectional	Seção transversal de curvatura
curv_general	Disposição geral
curv_longitudinal	Disposição longitudinal
Mass_balance_index	Índice de balanço de massa
normalized_height	Altura normalizada
slope_degress	Graus de inclinação
terrain_ruggedness_index	Índice de robustez do terreno
vector_ruggedness_index	índice de robustez vetorial

3.2 Covariáveis utilizadas com mais frequência pelos algoritmos

A partir da contagem das covariáveis mais utilizadas pelo RFE (Figura 5) observa-se que os algoritmos utilizaram diferentes variáveis explicativas para a

predição das classes de solo na área estudada. *Standardized height* (altura padronizada) e *Effective air flow heights* (altura efetiva do fluxo de ar) foram utilizadas pelos três algoritmos em quase 100% das repetições demonstrando grande relevância para o presente mapeamento.

Altura padronizada e altura efetiva do fluxo de ar são covariáveis que têm relação com a posição dos pontos amostrais em relação às cotas altimétricas (Tabela 2), indicam que grande parte das classes de solo são explicadas pela posição em que se encontram nas topossequências, enfatizando o grau de importância do relevo como fator de formação de solo.

O algoritmo C5, diferente dos outros modelos, utilizou com frequência as covariáveis longitude e latitude, isso se deve pela classificação da unidade de mapeamento PACdx + SNo, essas classes de solo aparecem somente a oeste do campus universitário, devido a uma combinação de covariáveis presentes nessas coordenadas, é uma região com grande número de nascentes, drenagem imperfeita e material de origem especifico (depósitos detrito lateríticos) ou seja, sem descontinuidade litológica.

A litologia foi importante para classificação feita pelo RF, a área estudada apresenta duas topossequências diferentes, a região sudoeste, em sua totalidade, está localizada sob depósitos detrito lateríticos tércio-quaternários ou sedimentos recém depositados às margens de córregos e rios que drenam a região. As outras áreas possuem sequências de solos, tanto originados a partir dos depósitos sedimentares, quanto diretamente de rochas do cristalino, expostas por processos erosivos. Esta sequência, além de apresentar solos na meia encosta, formados sob descontinuidade litológica também apresentam solos formados sob sedimentos recém depositados próximo aos cursos hídricos da universidade.

Para o RF e C5 o curso hídrico foi importante para delimitar as áreas de solos hidromórficos oriundos dos depósitos quaternários.

Além das feições mencionadas, o SVM levou em consideração a profundidade de vale e o Índice de multiresolução da planicidade das áreas de topo em todas as repetições, indicando que as características do relevo plano foram a maior fonte de informação do algoritmo para classificação e isso se deve pela grande

porção da classe Latossolo que encontramos nas áreas de maior altimetria e menor declividade.

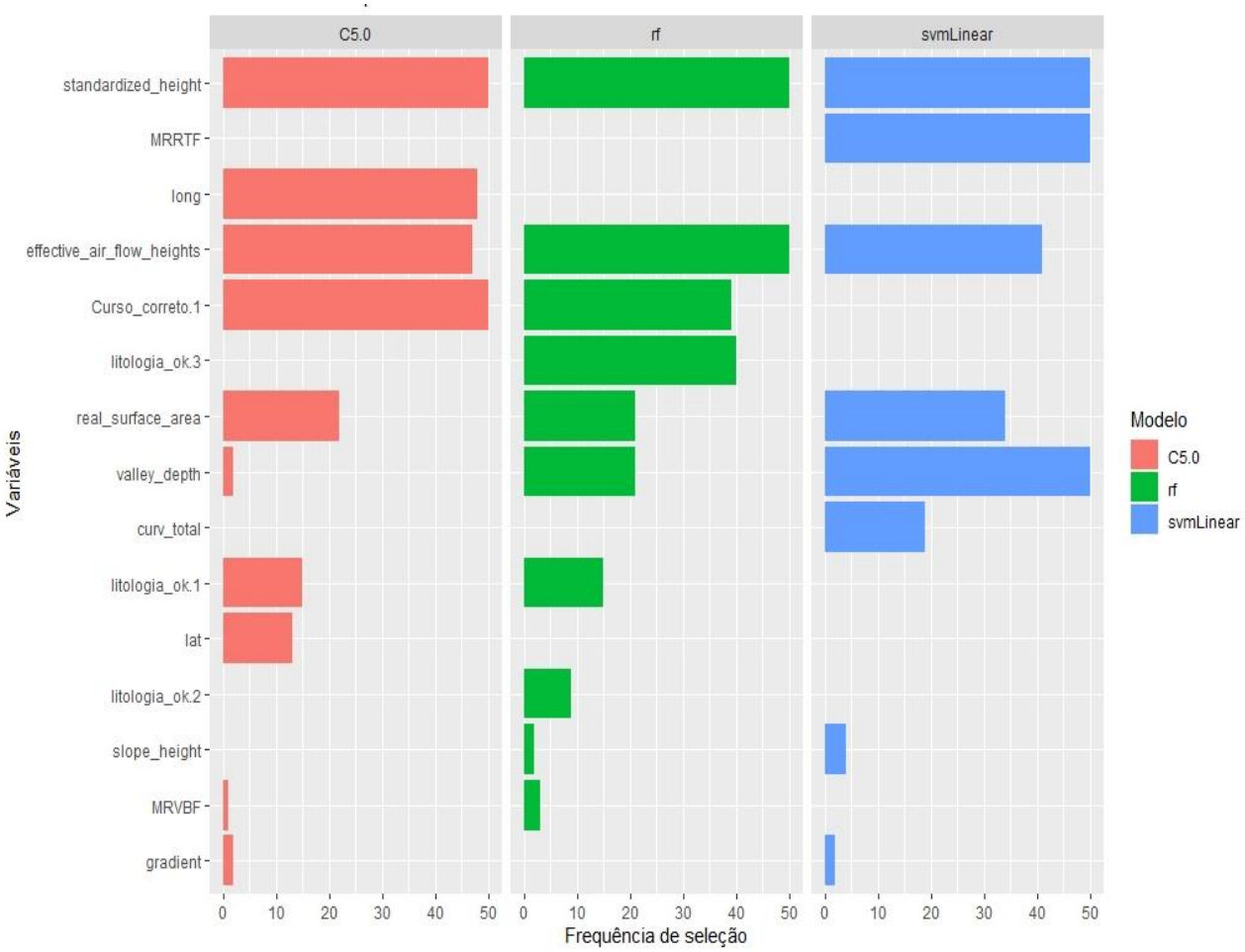

Figura 3 - Frequência da seleção de covariáveis feitas pelo RFE.

Tabela 2 - Covariáveis mais utilizadas pelos algoritmos para localizar as classes de solos ao longo da paisagem.

Código da covariável	Nome	Descrição
curso_correto	Curso hídrico	Áreas onde passam cursos hídricos
curv_total	Curvatura total	Medida geral da curvatura da superfície
effective_air_flow_heights	Alturas de fluxo de ar eficazes	Cálculo da que leva em consideração a distância entre picos e vales e correlaciona o resultado com a altura média da área de estudo
gradient	Gradiente de distância de descida	Cálculo de um novo índice topográfico para quantificar os controles de descida na drenagem local
lat	Latitude	Número referente a latitude em metros (coordenadas UTM)
litologia_ok	Material de origem	Delimitação dos pacotes de material de origem existentes no campus
long	Longitude	Número referente a longitude em metros (coordenadas UTM)
MRRTF (Multiresolution index of ridge top flatness)	Índice de multiresolução da planicidade do topo do cume	Indica posições planas em áreas de alta altitude
MRVBF (Multiresolution index of valley bottom flatness)	Índice multiresolução de planicidade do fundo do vale	Indica superfícies planas no fundo do vale
real_surface_area	Área de superfície real	Cálculo real da área da célula
slope_height	Altura do declive	Distância vertical entre a base e o cume da encosta
standardized_height	Altura padronizada	Distância vertical entre a base e o índice de inclinação padronizado
valley_depth	Profundidade do vale	Cálculo da distância vertical ao nível da base de drenagem

3.3 Performance dos algoritmos

Os valores de performances dos modelos apontam que o desempenho foi semelhante entre os algoritmos C5 e RF, enquanto os resultados de performance do SVM foram ligeiramente mais baixos. Utilizando os 50 resultados armazenados no quadro de dados, observou-se que a mediana da acurácia foi de 0,68 e 0,62 para os modelos C5 e RF, com desvio padrão de 0,07 e 0,08, respectivamente; enquanto que o SVM apresentou uma mediana de 0,56, com desvio padrão de 0,08 (Quadro 2).

Com relação ao índice Kappa (figura 5), a mediana foi de 0,64 para o C5 com desvio padrão de 0,07; para o RF 0,57 com variância de 0,08, para o SVM 0,51 com desvio padrão de 0,08 como mostra (Quadro 2).

Ao observar os gráficos violinos (Figura 3 e 4), as partes mais largas apontam maior volume de resultados referentes a performance de cada modelo. De acordo com Landis e Koch, (1977), os valores de Kappa para o algoritmo C5 a concordância é classificada como forte (entre 0,61 até 0,8), no caso do SVM e do RF a concordância é classificada como moderada (entre 0,41 e 0,6), a partir das variáveis selecionadas.

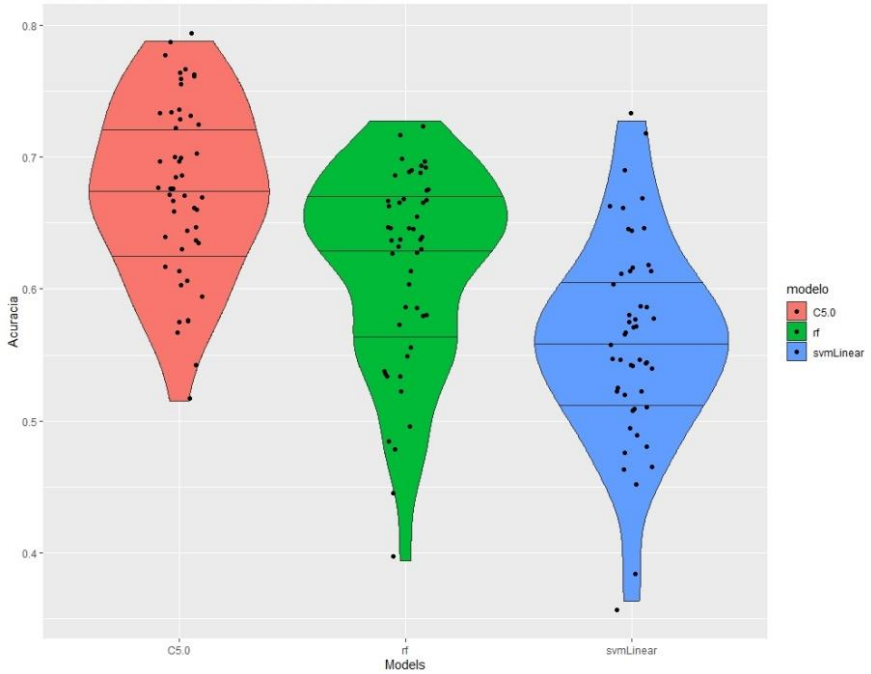

Figura 4 - Resultados da acurácia dos modelos RF (verde), C5 (vermelho) e SVM (azul) em um mapeamento digital de solos em Tabuleiros Interioranos.

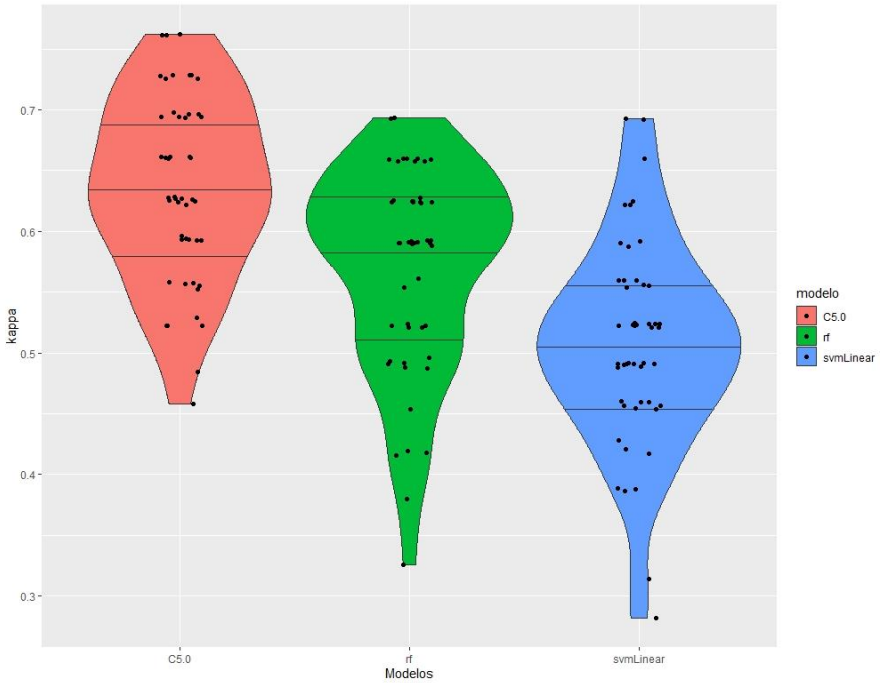

Figura 5 - Resultados do índice Kappa modelos RF (verde), C5 (vermelho) e SVM (azul) em um mapeamento digital de solos em Tabuleiros Interioranos.

Quadro 2 - Métricas de performance dos algoritmos C5, RF e SVM Kernel Linear para Mapeamento digital de solos em tabuleiros interioranos.

Modelo	Kappa	Desvio padrão	Coeficiente de variação
C5	0.64	0.07	11.78
RF	0.57	0.08	14.92
svmLinear	0.51	0.08	16.70
Modelo	Acurácia	Desvio padrão	Coeficiente de variação
C5	0.68	0.07	9.85
RF	0.62	0.08	12.26
svmLinear	0.56	0.08	13.47

A variância dos resultados deixa clara a importância de incluir repetição de rotinas de modelagem no mapeamento digital de solos, esse método evita superestimar ou subestimar o desempenho dos classificadores. Observando os resultados do presente estudo, hipoteticamente com uma única repetição, o valor de Kappa poderia ser classificado como razoável (entre 0,21 e 0,4) ou forte (entre 0,61 e 0,8) para todos os algoritmos, assim seria difícil identificar o real poder de predição e comparar os seus resultados para eleger o melhor.

Os resultados obtidos no presente trabalho são coerentes com outros encontrados por diferentes autores. Quando comparado aos resultados encontrados por Meier et al. (2018), de índice Kappa variando entre 0,42 e 0,48

em mapas de classe de solo gerados por diferentes algoritmos de ML (entre eles SVM e RF), em Minas Gerais - Brasil, o presente artigo apresenta resultados de ajustes ligeiramente superiores.

Utilizando o RF para um mapeamento digital em São Paulo, Mendes e Demattê (2022) encontraram acurácia geral de 0,86 e índice capa de 0,81 a nível de ordem de solos, o local de estudo abrange oito cidades e quase 2.598 km2, apesar de diferentes metodologias e nível de detalhamento os resultados apresentaram melhores performances do que as métricas encontradas no presente estudo, entretanto ambos os trabalhos confirmam o poder do algoritmo RF para predição de classes de solo no Brasil.

Este artigo reforça os resultados de Giasson et. al. (2011) em um mapeamento digital de solos em encostas basálticas subtropicais em Porto Alegre – Brasil onde os autores afirmam que o uso de árvores de decisão foi eficaz na predição de ocorrência de unidades de mapeamento de solos na área estudada.

Resultados muito semelhantes foram encontrados por Gonçalves et. al. (2021) utilizando algoritmo baseado em árvore de decisão, os pesquisadores encontraram 0,66 e 0,62 de índice Kappa e acurácia respectivamente, mapeando a distribuição de classes de solo em grandes áreas, baseando-se em mapas de solo preexistentes de áreas menores e semelhantes em uma região tropical no município de Itajubá, localizado no sul do estado de Minas Gerais, Brasil.

Abordando outras áreas ao redor do mundo, no noroeste do Irã, Sharififar A. et. al. (2019), comparando C5, RF e outros algoritmos de ML, também observaram que o C5 obteve maiores resultados de validação (Kappa 0,14).

Em um MDS localizado em uma área sedimentar no Oeste do Haiti, o algoritmo RF obteve performance de índice Kappa 0,64, assim como neste trabalho o resultado reflete sua capacidade para processar e generalizar dados de solo (JEUNE et al., 2018).

Na mesma linha, Zeraatpisheh M. et al. (2017) comparando algoritmos de ML, demonstrou que o classificador baseado em árvores de decisão apresentou resultados com maior desempenho (Kappa 0,33) nos níveis taxonômicos de grupo e subgrupo em um MDS localizado no semiárido do Irã.

Assim como ao comparar RF, C5, e outros modelos matemáticos para mapeamento de solos na região de Baneh no Irã, Taghizadeh-Mehrjardi R. et al. (2015) chegaram à conclusão que entre os algoritmos de arvores de decisão testados o C5 foi eficiente para prever famílias do solo, e concluem que o preditor é recomendado para gerar modelos de previsão espacial de solos.

Maleki S. et al (2020) estudando o efeito da precisão dos dados topográficos na melhoria do mapeamento digital do solo também no Irã, encontrou um índice Kappa de 0,23 na predição de solos a nível de família utilizando RF.

3.4 Mapas estatísticos derivados da modelagem com repetição

Os mapas de mudança (Figura 7) representam as áreas que foram identificadas pela mesma unidade de mapeamento nas 50 repetições de cada modelo, evidenciando áreas da paisagem em que os algoritmos encontram um padrão de covariáveis explicativas que são associadas a alguma unidade de mapeamento especifica. As áreas de maior altimetria e baixa declividade (os topos planos dos tabuleiros e as baixadas, associadas às linhas de drenagem, respectivamente, tiveram classificação unânime quanto a unidade de mapeamento pelos algoritmos RF e C5. Isso se deve a maior homogeneidade dos materiais geológicos e das classes de relevo que que dão origem a solos com pequena variabilidade de classes nessas áreas. Nos topos planos de tabuleiro e no terço superior das encostas onde, por apresentarem classes de relevo plano a suave ondulado, ocorre pouco transporte de material e a água tem a tendência de percolar, existe a formação de solos profundos e bem intemperizados com predominância dos Latossolos Amarelos. Nas áreas de baixadas planas, associadas as linhas de drenagem, ocorrem solos hidromórficos tais como os Gleissolos.

Estas áreas específicas representam 16%, 20% e 27% da área total mapeadas pelo SVM, C5 e RF respectivamente, apresentam domínios de solos (Latossolos Amarelos e Gleissolos Háplicos), confirmados por diferentes levantamentos (SOUZA & SOUZA, 2001; RIBEIRO et al, 1998; RIBEIRO et al, 1988; RODRIGUES et al, 2003) realizados na região e expressos em mapas de solos em diferentes escalas.

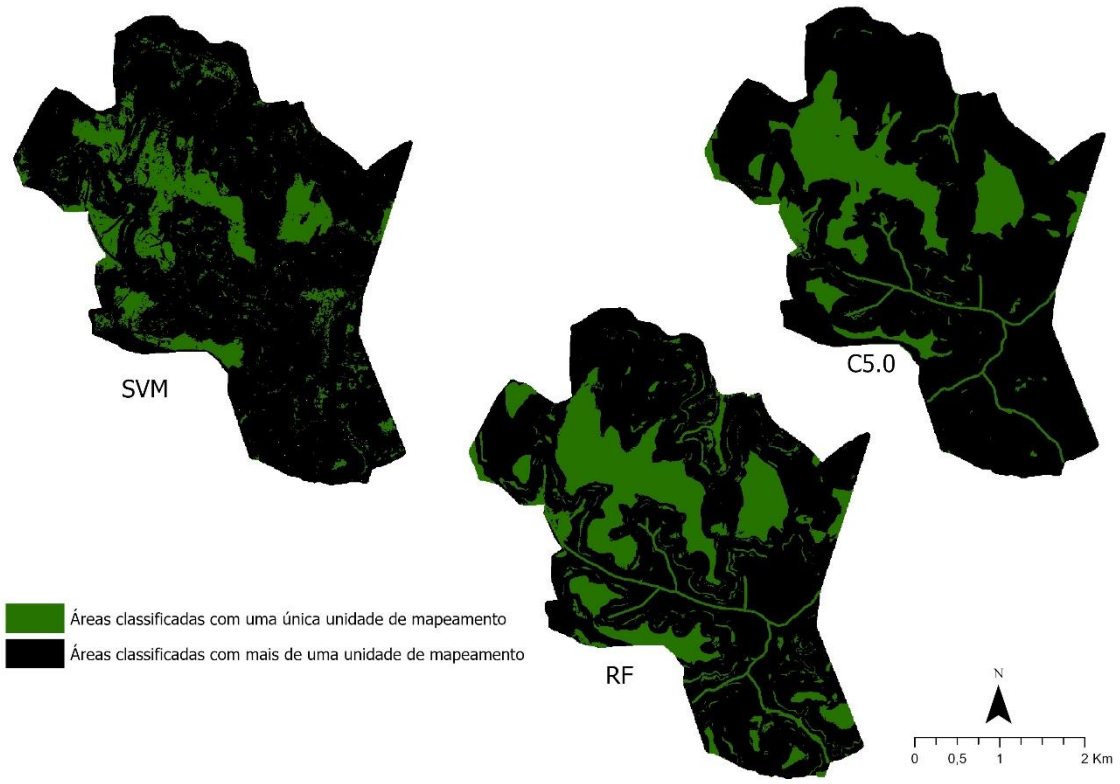

Figura 6 - Mapas das áreas classificadas apenas com uma unidade de mapeamento de classes de solos pixel a pixel em área de Tabuleiro Interiorano no Recôncavo da Bahia pelos algoritmos Random Forest (RF), Árvore de decisão C5.0 (C5) e Maquina de vetores suporte.

As áreas ocupadas com mais de uma unidade de mapeamento, representam os ambientes de terços médio e inferior de encosta e baixada, suave ondulada e plana. A grande variabilidade de solos nestes ambientes estão associadas a diversidade de materiais de origem, que em alguns casos, estão presentes no mesmo solo (descontinuidade litológica), observadas em Cambissolos Háplicos distróficos; à espessura variada do pacote sedimentar (depósitos detrítico laterítico), que recobre as rochas do embasamento cristalino; aos diferentes graus de dissecação dos Tabuleiros, por processos erosivos; e aos diferentes padrões de sedimentação de material recente que recobrem ou não as rochas cristalinas expostas nestes ambientes.

Meier (2018) relata que as covariáveis utilizadas em seu estudo apresentaram dificuldades na distinção de alguns solos em terrenos dissecados enfatizando um dos desafios do MDS. Comparando os mapas de mudança de classe de solo é possível perceber que o algoritmo RF teve maior área onde a classificação foi de apenas uma unidade de mapeamento, além dos topos planos e da linha de drenagem, algumas áreas de encosta também não sofreram

mudança nas repetições. O SVM foi o algoritmo que demonstrou menor área classificada com apenas uma unidade de mapeamento.

Os mapas de variância da unidade de mapeamento (Figura 8) reforçam o fato que as encostas mais escarpadas foram mais difíceis de serem classificadas quando comparadas com as áreas mais planas, essas regiões tiveram até 6 UMs (cores laranjas) diferentes atribuídas aos seus pixels. No caso do algoritmo C5, essas áreas foram menores. Pequenas regiões localizadas a noroeste e nordeste (áreas circuladas em vermelho) foram as mais complexas para classificação, essas regiões receberam entre 7 e 8 unidades de mapeamento diferentes de 10 disponíveis no caso do C5, essas duas áreas estão às margens da área estudada, onde existe comunidades urbanas que se fazem presente (Baixa da linha e Sapucaia, respectivamente). O desenvolvimento urbano observado acaba influenciando de forma significativa o MDS, principalmente, por conta das construções civis e distribuição heterogênea da vegetação. O algoritmo Random Forest também apresentou incerteza de classificação em uma região ao sul (circulada em vermelho) essa área contem solos encharcados, uma vegetação em estágio de regeneração primaria com algumas árvores presentes e um relevo suave ondulado.

O mapa com áreas de maior incerteza foi o mapa gerado pelo SVM, seguido pelo RF e com o melhor resultado o C5.

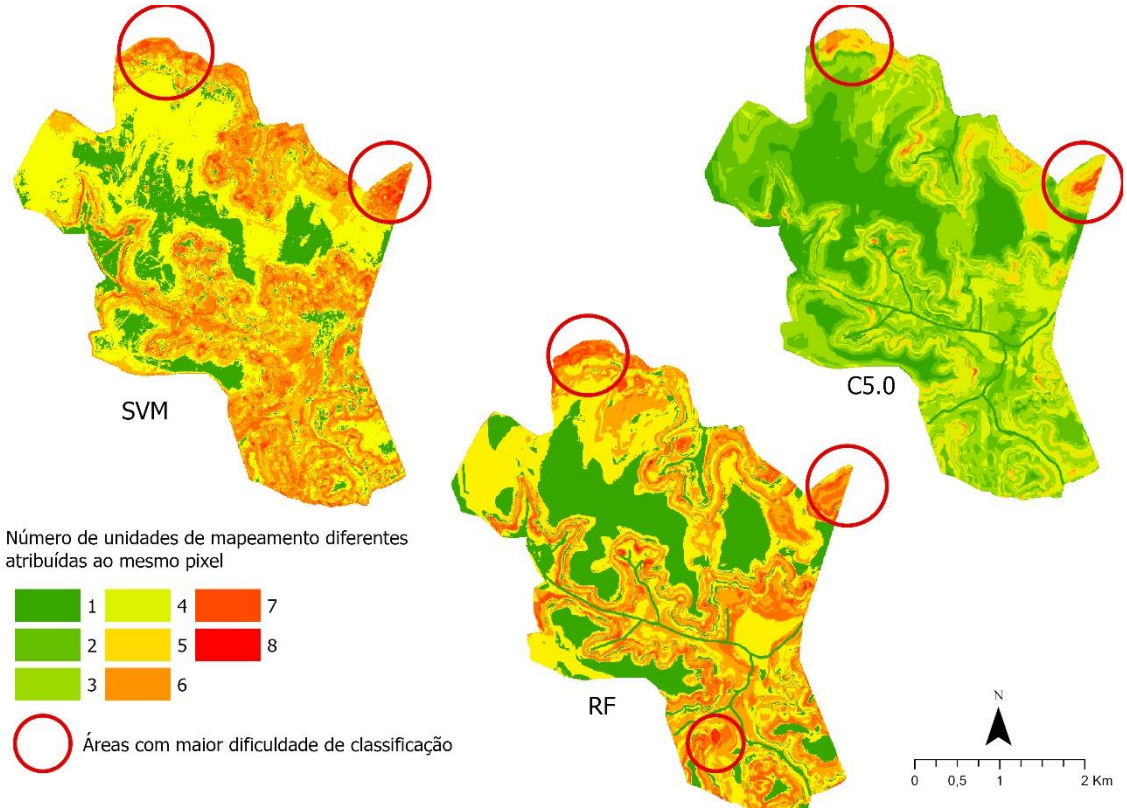

Figura 7 - Mapas de variância de pixel a pixel de unidades de mapeamento de solos em área de Tabuleiro Interiorano do Recôncavo da Bahia utilizando aprendizado de máquina, Randon Forest (RF), Árvore de decisão C5.0 (C5) e Maquina de vetores suporte Kernel Linear.

Os mapas de frequência da moda (Figura 9), completam a informação dos mapas anteriores, nele é possível identificar o grau da incerteza em cada local, as áreas vermelhas indicam que a moda foi atribuída 12 vezes ao pixel, o mínimo para estabelecer uma moda com o conjunto de dados seria de 6 vezes. Esses locais onde a moda foi atribuída poucas vezes (áreas circuladas em vermelho) são locais prioritários para realizar novas amostragens, em um futuro mapeamento da área.

A frequência da moda do RF mostra que a maior parte da área de estudo teve a grande representatividade dá UM modal (Latossolos) nas repetições, esse algoritmo demonstrou maior constância em classificar os padrões quando comparada com os outros dois classificadores. O C5 demonstra grande capacidade de encontrar padrões na distribuição das classes de solo da área, porem na região noroeste, como já comentada anteriormente, houve uma acentuada incerteza.

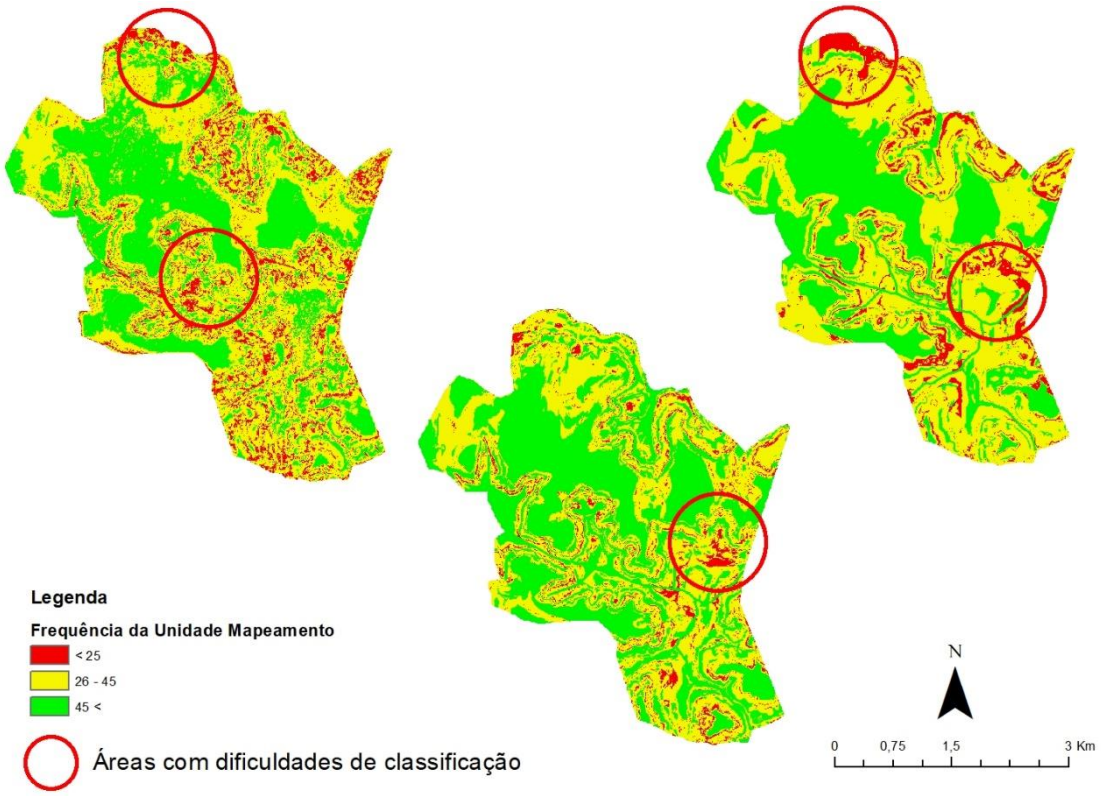

Figura 8 - Mapas de frequência da moda pixel a pixel de unidades de mapeamento de solos em área de Tabuleiro Interiorano do Recôncavo da Bahia utilizando aprendizado de máquina, Random Forest (RF), Árvore de decisão C5.0 (C5) e Maquina de vetores suporte.

Os mapas da distribuição da moda (Figura 10) foram obtidos selecionando a UM modal de 50 repetições para cada pixel, esses mapas demonstram concordância com os levantamentos anteriores para os algoritmos RF e C5. É possível também identificar dificuldades encontradas por esses dois algoritmos, para o RF na região nordeste, e para o C5 na região noroeste, que foi classificado como a unidade de mapeamento PACdx, em ambos casos, o que não condiz com a realidade de campo. A falta de pontos de observação nas duas regiões, assim como a urbanização dessas áreas (Sapucaia e Vila Guaxinim, respectivamente) influenciaram no treinamento dos modelos.

A falta de pontos de observação reflete em uma menor variedade de informação das covariáveis explicativas, a urbanização (construções civis) confunde os padrões ambientais (como declividade, altitude e refletância registrada pelos sensores nos satélites) tornando essas áreas de difícil modelagem, outro aspecto importante a se observar é que como foi utilizado um MDS, a vegetação teve influência no grau de declividade das encostas tornando

elas mais suaves e favorecendo extrapolação das áreas referentes ao Planossolo na base das encostas.

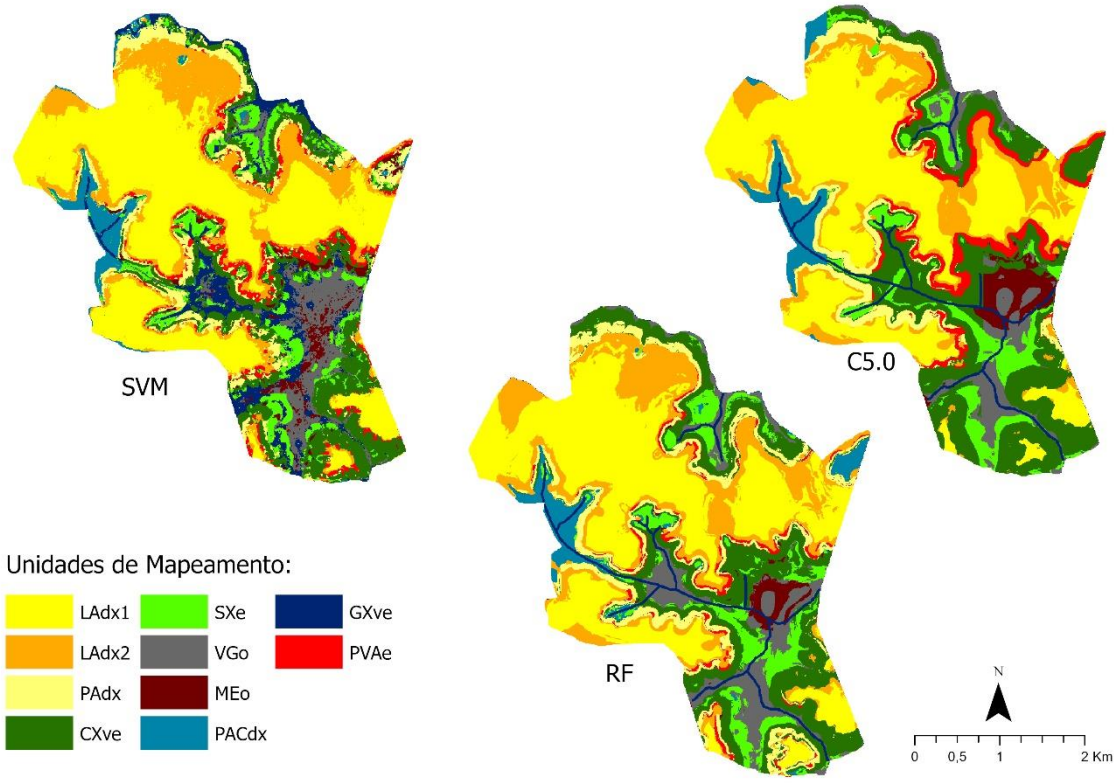

Figura 9 - Mapas da moda pixel a pixel de unidades de mapeamento de solos em área de Tabuleiro Interiorano do Recôncavo da Bahia utilizando aprendizado de máquina, Random Forest (RF), Árvore de decisão C5.0 (C5) e Maquina de vetores suporte Kernel Linear (SVM).

A classificação das áreas referentes ao Chernossolo foi bastante condizente com a realidade de campo. São áreas da paisagem onde, o cristalino aflora e ainda existe grande parte da vegetação nativa, em pequenas partes das encostas, e, de forma mais acentuada, em uma ilha de vegetação formada na baixada plana, próximo ao curso hídrico de maior volume da área.

Utilizando o mapa da moda do C5 (de maior desempenho) é possível identificar que o Latossolo Amarelo é predominante na área de estudo, em seguida do Cambissolo Vértico (Quadro 1).

Quadro 3 - Distribuição das Unidades de Mapeamento do mapa da moda do algoritmo C5

UM	Ha	%
LAdx1	450.8	32.98
LAdx2	189.0	13.83
PAdx	85.8	6.28
CXve	265.2	19.40
SXe	93.6	6.85
VGo	77.6	5.68
MEo	39.6	2.90
PACd	51.3	3.75
GXv	36.2	2.65
PVAe	77.9	5.70
Total	1367	100

As menores manchas são as do Gleissolo e do Chernossolo, 2.49% e 2.72% da área respectivamente.

4. CONCLUSÃO

Com o conjunto de dados utilizado, o classificador C5 obteve melhores índices de desempenho em predição de classes de solo, seguido do algoritmo RF, em um mapeamento digital realizado na região dos tabuleiros interioranos (região tropical do Brasil). por apresentarem Índice Kappa e parecidos. Enquanto o modelo SVM teve performance menor, porem também demonstrou potencial para classificação de solos na área de estudo.

A variação dos resultados de índices Kappa e Acurácia indicam que utilizar repetição variando os conjuntos teste e treino é uma boa prática para avaliar algoritmos no mapeamento digital de solos.

As covariáveis com maior potencial de explicar a distribuição das unidades de mapeamento na região, independentemente do algoritmo testado, são Standardized height (altura padronizada) e effective air flow heights (altura efetiva do fluxo de ar).

A distribuição das classes de solo de solo do campus da universidade é marcada pela predominância dos Latossolos Amarelos nos topos planos dos tabuleiros e à medida que o terreno começa a ficar ondulado outras classes se fazem presentes como é o caso do Cambissolo, Argissolo, Planossolo e Chernossolo, próximo à linha de drenagem encontramos o Vertissolo e o Gleissolo.

5. REFERÊNCIAS

BAGATINI, T.; GIASSON, E.; TESKE, R. Seleção de densidade de amostragem com base em dados de áreas já mapeadas para treinamento de modelos de árvore de decisão no mapeamento digital de solos. **Revista Brasileira de Ciência do Solo**, v. 39, n. 4, p. 960–967, 1 jul. 2015.

BREIMAN L. **Random forests. Mach Learn**. 2001; 45:5-32. https://doi.org/10.1023/A:1010933404324

BREIMAN L. randomForest: Breiman and Cutler's Random Forests for Classificationand Regression. **R package version 1.2.0**; 2006.

BRENNING A.; BANGS D.; BECKER B.; SCHRATZ P.; POLAKOWSKI F.; RSAGA: SAGA Geoprocessing and terrain analysis**. R package version 1.2.0**; 2018. Available from: https://CRAN.R-project.org/ package=RSAGA

CONGALTON RG.; GREEN K.; Assessing the accuracy of remotely sensed data: principles and practices. **2nd ed. Boca Raton: CRC Press**; 2009.

CORTES C.; VAPNIK V.; 1995. Support-vector networks. Mach. Learn. 20 (3), 273–297. **R package version 1.2.0**; 2018

CPRM – SERVIÇO GEOLÓGICO DO BRASIL. **Mapa Geológico do estado da Bahia**. [Rio e Janeiro], 2001. mapa, color., 72cm x 85,5cm. Escala: 1:1.000.000.

DIRETORIA DE SERVIÇO GEOGRÁFICO (DSG). **Banco de Dados Geográficos do Exército. Versão 3.0**. 2013. Disponível em: <http://www.geoportal.eb.mil.br/mediador/>. Acesso em: 05 abril 2020.

GIASSON E.; SARMENTO E.; WEBER E.; FLORES C.; HASENACK H.; Decision trees for digital soil mapping on subtropical basaltic steeplands. **Sci. Agric. (Piracicaba, Braz.),** v.68, n.2, p.167-174, March/April 2011.

GOMES L. C.; FARIA R.; SOUZA E.; VELOSO G.; SCHAEFER C.; FILHO E.; Modelling and mapping soil organic carbon stocks in Brazil, **Geoderma,** Volume

340, 2019, Pages 337-350, ISSN 0016-7061, https://doi.org/10.1016/j.geoderma.2019.01.007.

GONÇALVES T.; PONS N.; MELLONI E; CURI M. Digital soil mapping: Predicting soil classes distribution in large areas based on existing soil maps from similar small areas. **Ciência e Agrotecnologia**. 2021.

GUO, Z. Landslide susceptibility zonation method based on C5.0 decision tree and K-means cluster algorithms to improve the efficiency of risk management. **Geoscience Frontiers**, v. 12, n. 6, p. 101249, 1 nov. 2021.

JENNY H. Fatores de formação do solo, um sistema de pedologia quantitativa. **McGraw-Hill**, Nova York 1941.

JEUNE, WESLY Regressão logística multinomial e classificadores florestais aleatórios no mapeamento digital de classes de solo no oeste do Haiti. **Rev. Bras. Ciênc. Solo**, Viçosa, v. 42, e0170133, 2018.

KARATZOGLOU A.; SMOLA A.; HORNIK K.; kernlab: kernel-based machine learning lab**. R package version 0.9-25**; 2016. Available from: https://CRAN.R-project.org/package=kernlab

KEMPEN, B. Efficiency Comparison of Conventional and Digital Soil Mapping for Updating Soil Maps. **Soil Science Society of America Journal**, v. 76, n. 6, p. 2097–2115, 1 nov. 2012.

KHALEDIAN Y.; MILLER B. A. Selecting appropriate machine learning methods for digital soil mapping. **Applied Mathematical Modelling**, v. 81, p. 401–418, 2020.

KUHN M. Caret: classification and regression training. **R package version 6.0-76; 2017**. Available from: https://CRAN.R-project.org/package=caret.

KUHN M. Predictive modeling with R and the caret Package. **Google Scholar**; 2013.

KUHN, M., WESTON, S., CULP, M., COULTER, N., QUINLAN. Package 'C50.' **R Package**, 2018.

LANDIS JR.; KOCH GG. The measurement of observer agreement for categorical data. **Biometrics**. 1977; 33:159-74. https://doi.org/10.2307/2529310

LIEß, M. At the interface between domain knowledge and statistical sampling theory: Conditional distribution based sampling for environmental survey (CODIBAS). **CATENA**, v. 187, p. 104423, 1 abr. 2020.

MALEKI S.; KHORMALI F.; JAHANGIR M.; PATRICK B.; BODAGHABADI M., Effect of the accuracy of topographic data on improving digital soil mapping predictions with limited soil data: An application to the Iranian loess plateau, **CATENA**, Volume 195, 2020, 104810, ISSN 0341-8162, https://doi.org/10.1016/j.catena.2020.104810.

MALONE BRENDAN P. Using R for digital soil mapping. Basel, **Switzerland: Springer International Publishing**, 2017.

MASCARENHAS J.; PEDREIRA A.J.; GIL C.; NEVES P.; OLIVEIRA, J. E. DE & MARINHO M.M. Geologia da região centro oriental da Bahia. Projetos Bahia-Bahia II -Sul da Bahia. Relatório integrado. Brasília, **Ministério de Minas e Energia. Departamento Nacional da produção Mineral**, 1979, 128 p.

MCBRATNEY A.B.; MENDONÇA SANTOS M.L.; MINASNY B.; On digital soil mapping, **Geoderma**, Volume 117, Issues 1–2, 2003, Pages 3-52, ISSN 0016-7061, https://doi.org/10.1016/S0016-7061(03)00223-4.

MEIER M.; SOUZA E.; FRANCELINO M.R.; FERNANDES FILHO E.I.; SCHAEFER, C. 2018. Mapeamento digital de solos usando algoritmos de aprendizado de máquina em uma área tropical montanhosa. **Revista Brasileira de Ciência do Solo** 42: 1 22.

MENDES W.; DEMATTÊ J. Digital soil mapping outputs on soil classification and sugarcane production in Brazil. **Journal of South American Earth Sciences**. **Elsevier**. 2022.

MINASNY B.; MCBRATNEY A. A conditioned Latin hypercube method for sampling in the presence of ancillary information. **Computers & Geosciences**, v. 32, n. 9, p. 1378–1388, 1 nov. 2006.

MURUGAN BHAGAVATHI S.; THAVASIMUTHU A.; MURUGESAN A. Weather forecasting and prediction using hybrid C5.0 machine learning algorithm**. Int J Commun Syst**. 2021; 34: e4805. https://doi.org/10.1002/dac.4805

QUINLAN J.R., 1993. C4. 5: Programming for machine learning. In: **Kauffmann, Morgan (Ed.)**, 38. pp. 48.

R Development Core Team. R: A language and environment for statistical computing. **R Foundation for Statistical Computing**. Vienna, Austria; 2016. Available from: http://www.R-project.org/.

RADAMBRASIL – Levantamento de recursos naturais. Vol. 24. Folha SD.24-Salvador: Geologia, Geomorfologia Pedologia, Vegetação, Uso potencial da Terra. Rio de Janeiro, **Ministério das Minas e Energia**. Projeto Radam Brasil. 1981. 620p.

RIBEIRO, L. P. Levantamento dos solos, capacidade de uso das terras e classificação de terras para irrigação da estação de plasticultura da UFBa. / Politeno. Campus da Escola de Agronomia – Cruz das Almas – Ba. **Universidade Federal da Bahia.** Salvador – Bahia 1991

RIBEIRO, L. P. Levantamento dos solos, capacidade de uso das terras e classificação de terras para irrigação da área do Candeal - Bahia. Campus da Escola de Agronomia – Cruz das Almas – Ba. **Universidade federal da Bahia.** Salvador – Bahia 1988

SHARIFIFAR A., FEREYDOON SARMADIAN, BRENDAN P. MALONE, BUDIMAN MINASNY, Addressing the issue of digital mapping of soil classes with imbalanced class observations, **Geoderma**, Volume 350, 2019, Pages 84-92, ISSN 0016-7061, https://doi.org/10.1016/j.geoderma.2019.05.016.

SILVA, C. Diferentes classificadores na predição de classes de solos em mapeamento digital. Anais XVI Simpósio Brasileiro de Sensoriamento Remoto - **SBSR**, 2011.

SODRÉ L. Maquete e macromonolitos como ferramenta de auxilio didático no ensino de solos. Trabalho de conclusão de curso submetido ao Colegiado de Graduação de Tecnologia em Agroecologia do Centro de Ciências Agrárias, Ambientais e Biológicas da **Universidade Federal do Recôncavo da Bahia**, 2019

SOUZA L.; SOUZA L. D. Caracterização físico-hídrica de solos da área do Centro Nacional de Pesquisa de Mandioca e Fruticultura Tropical. Boletim de Pesquisa e Desenvolvimento nº 20 **EMBRAPA**. Dezembro 2001

TAGHIZADEH-MEHRJARDI R.; NABIOLLAHI K.; MINASNY B.; TRIANTAFILIS J. Comparing data mining classifiers to predict spatial distribution of USDA-family soil groups in Baneh region, Iran, **Geoderma**, Volumes 253–254, 2015, Pages 67-77, ISSN 0016-7061, https://doi.org/10.1016/j.geoderma.2015.04.008.

TEN CATEN A. Mapeamento digital de classes de solos: características da abordagem brasileira. **Ciência Rural**, v. 42, n. 11, p. 1989–1997, nov. 2012.

WADOUX A-C.; SAMUEL-ROSA A.; POGGIO L.; MULDER VL. A note on knowledge discovery and machine learning in digital soil mapping. **Eur J Soil Sci.** 2020; 71: 133– 136. https://doi.org/10.1111/ejss.12909