Universidade de São Paulo
Escola Superior de Agricultura “Luiz de Queiroz”
Centro de Energia Nuclear na Agricultura

Potencial agrícola no Cerrado utilizando ferramentas de aprendizado de máquina

Mariane Cristina do Amaral Romeiro

Dissertação apresentada para obtenção do título de Mestra em Ciências. Área de concentração: Ecologia Aplicada

Piracicaba
2022

Mariane Cristina do Amaral Romeiro
Engenheira Agrônoma

Potencial agrícola no Cerrado utilizando ferramentas de aprendizado de máquina

versão revisada de acordo com a resolução CoPGr 6018 de 2011

Orientador:
Prof. Dr. **GERD SPAROVEK**

Dissertação apresentada para obtenção do título de Mestra em Ciências. Área de concentração: Ecologia Aplicada

Piracicaba
2022

Dados Internacionais de Catalogação na Publicação
DIVISÃO DE BIBLIOTECA – DIBD/ESALQ/USP

Romeiro, Mariane Cristina do Amaral

Potencial agrícola no Cerrado utilizando ferramentas de aprendizado de máquina / Mariane Cristina do Amaral Romeiro. - - versão revisada de acordo com a resolução CoPGr 6018 de 2011. - - Piracicaba, 2022.

52 p.

Dissertação (Mestrado) - - USP / Escola Superior de Agricultura “Luiz de Queiroz”. Centro de Energia Nuclear na Agricultura

1. Zoneamento agrícola 2. Aprendizado de máquina 3. *Random forest* 4. Cerrado I. Título

AGRADECIMENTOS

Agradeço à minha mãe, mulher batalhadora que sempre me incentivou a estudar e fez todo o possível para que isto se realizasse. Agradeço também à minha avó, que já não está nesse mundo, mas que sempre me ensinou valiosas lições sobre a vida.

Agradeço ao meu companheiro Paulo pela compreensão e carinho ao longo da elaboração deste estudo. Aos meus amigos de Piracicaba e de São Paulo que sempre me incentivaram e me auxiliarem sempre que precisei, em especial o Gustavo, Gislaine, Karine, Stella e Tatiane, vocês são muito importantes para mim!

Ao meu orientador, Prof. Dr. Gerd Sparovek, por ter me aceitado como orientada, pelas boas conversas, ideias e aprendizados que tivemos. Obrigada pela paciência e disposição em me ajudar ao longo do trabalho.

À Agroicone, empresa que trabalho há 6 anos, pelo incentivo em fazer o mestrado e dar mais esse passo em minha carreira. Em especial a Leila, minha supervisora direta, que sempre foi muita humana e me deu todo apoio.

A todos os professores e profissionais da ESALQ/USP que de alguma forma me auxiliaram durante toda a jornada.

SUMÁRIO

RESUMO

Potencial agrícola no Cerrado utilizando ferramentas de aprendizado de máquina

O planejamento territorial é uma ferramenta de suma importância para o desenvolvimento sustentável do setor agrícola brasileiro, ainda mais em biomas com fronteiras agrícolas em pleno crescimento, como o Cerrado. O zoneamento agrícola é um dos principais instrumentos do planejamento territorial, que em sua maioria é realizado a partir de análises multicritérios e depende de interpretações de analistas. Ao buscar por diferentes alternativas para este tipo de análise, veio o termo Inteligência Artificial (IA), ramo da ciência da computação, que vem sendo usado de forma abrangente não só no meio acadêmico, mas em funcionalidades usadas no dia a dia, como *streaming* de filmes e séries, carros inteligentes, reconhecimento facial, comportamento de consumo, entre muitos outros. Diante de tamanha versatilidade deste ramo, foi proposto para este estudo utilizar ferramentas de aprendizado de máquina, que é um dos campos da IA, para desenvolver modelos preditivos para classificação do potencial agrícola no Cerrado brasileiro. Foram desenvolvidos quatro modelos, utilizando o algoritmo *Random Forest*, entre eles o potencial de expansão da soja, potencial de intensificação da pecuária, potencial de expansão da floresta plantada e um modelo final que reúne as três cadeias agropecuárias. Como variáveis de entrada foram usados dados climáticos, edáficos, de infraestrutura e socioeconômicos, já como classe para treinamento do modelo utilizou-se dados de um estudo elaborado pela organização WWF-Brasil em parceria com outras instituições. A performance dos modelos foi avaliada a partir da matriz de confusão, e a melhor acurácia foi a do modelo de floresta plantada com 98%, seguido do modelo de soja e pecuária, com 86% e 79%, respectivamente. O modelo final apresentou uma acurácia geral de 80%.

Palavras-chave: Zoneamento agrícola, Aprendizado de máquina, *Random forest*, Cerrado

ABSTRACT

Agricultural potential in the Cerrado using machine learning tools

Territorial planning is an extremely important tool for the sustainable development of the Brazilian agricultural sector, even more so in biomes with fast growing agricultural frontiers, such as the Cerrado. The agricultural zoning is one of the main instruments of territorial planning, which is mostly carried out based on multi-criteria analysis and depends on analysts' interpretations. When searching for different alternatives for this type of analysis, the term Artificial Intelligence (AI) came up, a branch of computer science that has been widely used not only in the academic environment, but in functionalities used in everyday life, such as streaming movies and series, smart cars, facial recognition, consumer behavior, among many other features. Given such versatility of this branch, it was proposed for this study to use machine learning tools, which is one of the fields of AI, to develop predictive models for classification of agricultural potential in the Brazilian Cerrado. Four models were developed, using the Random Forest algorithm, among them the potential for soybean expansion, potential for livestock intensification, potential for planted forest expansion, and a final model that brings together the three agricultural chains. Climate, edaphic, infrastructure and socioeconomic data were used as inputs variables, while data from a study prepared by WWF-Brazil in partnership with other institutions was used as a class to train the model. The performance of the models was evaluated from the confusion matrix, and the best accuracy was the planted forest model with 98%, followed by the soybean and cattle ranching models, with 86% and 79%, respectively. The final model presented an overall accuracy of 80%.

Keywords: Agricultural zoning, Machine learning, Random forest, Cerrado

LISTA DE FIGURAS

LISTA DE TABELAS

1. INTRODUÇÃO

O Cerrado é o segundo maior bioma brasileiro, ocupando uma área de 198,3 milhões de hectares, o que representa 23,3% do território do país (IBGE, 2019). Devido a sua posição geográfica e às suas características ecológicas, o Cerrado tem importância fundamental para a sociedade brasileira tanto em termos de biodiversidade e manutenção dos recursos naturais, quanto relacionado à produção agrícola que se desenvolve no seu território.

O bioma é um grande pilar da agropecuária do Brasil, tanto que em 2015 foi responsável por 41% do valor total de produção das culturas temporária e perenes do Brasil (BOLFE; SANO; CAMPOS, 2020). Além do mais, é conhecido como "berço das águas" ou "caixa d'água do Brasil" pois nele encontram-se as nascentes das três maiores bacias hidrográficas da América do Sul: Amazônica/Tocantins, São Francisco e Paraná, o que resulta em um elevado potencial aquífero representado pelos aquíferos Urucuia, Bambuí e Guarani.

Os dados mais completos e atualizados sobre o uso da terra e ocupação do solo são do "MapBiomas - Collection 5" (2020). De acordo com essa fonte, em 2019, o Cerrado tinha uma área de 89,2 milhões de hectares (46,5%) com formação florestal, que inclui florestas, savanas e mangues. A atividade agrícola ocupava 86,9 milhões de hectares (43,8%), dos quais 25,9 milhões são para agricultura e 61 milhões, para pastagens, sendo que 23,7 milhões de hectares (39%) possuem algum tipo de degradação (LAPIG, 2019). Ainda assim, entre 2015 e 2020 ocorreu um avanço de 4,2 milhões de hectares da agropecuária sobre vegetação nativa ("Mapbiomas - Coleção 6", 2021)

De acordo com a FIESP (2020), há uma previsão de crescimento de 23,5% na área plantada de soja no Brasil, entre 2019 e 2029. A produção de carne deverá ter um crescimento de 23,3% no mesmo período. SPAROVEK *et al.*. (2011) afirma que com a intensificação da pecuária, é possível liberar áreas de pastagem com média e alta aptidão para expansão agrícola, sendo que um terço destas áreas estão no Cerrado.

O panorama mostra boas oportunidades no Cerrado. Para que sejam bem aproveitadas é fundamental planejamento territorial aliado com políticas públicas, a fim de alocar a expansão agrícola de forma sustentável. Um dos principais instrumentos de planejamento territorial é o zoneamento agrícola, muito difundidos no Brasil, como é o caso do Zoneamento Agrícola de Risco Climático (ZARC).

Trabalhando como pesquisadora na consultoria Agroicone desde 2016, foi possível desenvolver alguns estudos relacionados ao planejamento territorial do Cerrado. No mesmo ano, foi lançado uma análise sobre a dinâmica da expansão da soja no Cerrado (CARNEIRO FILHO *et al.*, 2018). Já em 2018, foram lançados mais dois estudos, sobre a oportunidade de expansão da soja no Cerrado (ROMEIRO *et al.*, 2018) e as áreas prioritárias para conservação no Cerrado (CARNEIRO FILHO *et al.*, 2018). Mais recentemente, em 2020, foi realizada uma série de webinars em parceria com o Grupo de Trabalho de Pastagens (GT-Pastagem) - formado pelas instituições WWF, TNC, Imaflora, Agroicone e Lapig - no qual foram apresentados os resultados de um estudo[1] sobre oportunidades de expansão agrícola em áreas de pastagens degradadas.

Diante da experiência com a temática no Cerrado, foram buscadas alternativas para avaliar o potencial de uso do solo, como a utilização de ferramentas de aprendizado de máquina, que possibilitam manipular grande quantidade de dados, são de fácil acesso e geram bons resultados. A partir desta busca foram encontrados alguns trabalhos como de LORENSINI (2019) que concluiu ser possível usar ferramentas de aprendizado de máquina para

[1] A apresentação do estudo está disponível em: https://www.agroicone.com.br/agroicone-apresenta-estudo-preliminar-em-webinar-gt-pastagens/

desenvolver modelo preditivo de aptidão agrícola para a região do Matopiba. Também foi possível gerar bons modelos para realizar predições de produtividade da cultura da soja em municípios brasileiros (GUIMARÃES, 2019).

E, juntando as particularidades do Cerrado com a vontade de aprender novas tecnologias, mais as referências de trabalhos anteriores, que este estudo foi desenvolvido.

2. OBJETIVOS

2.1 Objetivo geral

Desenvolver modelos preditivos para identificação do potencial agrícola das cadeias da soja, pecuária e floresta plantada para o Cerrado, utilizando aprendizado de máquina.

Avaliar a possibilidade de utilizar os modelos como ferramenta de planejamento territorial, buscando saber onde as áreas potenciais estão localizadas, qual o risco de abertura de novas áreas e a possibilidade de usar áreas abertas e/ou improdutivas para alocar a expansão das cadeias agropecuária.

2.2 Objetivos específicos

- Obter, organizar e padronizar os dados climáticos, edáficos, de infraestrutura e socioeconômicos, utilizando sistema de informação geográfica (SIG). Estes são os dados de entrada do modelo;
- Preparar os dados de treinamento do modelo baseado em estudos sobre o tema;
- Fazer o pré-processamento dos dados, escolher os melhores parâmetros e avaliar a performance dos modelos;
- Aplicar os modelos à base de dados completa e comparar com os dados originais para avaliar suas acurácias.

3. REVISÃO DE LITERATURA

A agricultura e pecuária são atividades muito importantes na economia do Brasil, o agronegócio representou 21,4% do Produto Interno Bruto (PIB) brasileiro em 2019, de acordo com cálculos realizados pelo Centros de Estudos Avançados em Economia Aplicada (Cepea), da Escola Superior de Agricultura Luiz de Queiroz – Universidade de São Paulo (ESALQ – USP), em parceria com a Confederação da Agricultura e Pecuária do Brasil (CNA) e com a Fundação de Estudos Agrários Luiz de Queiroz (Fealq).

A agricultura é uma atividade que depende das condições ambientais, principalmente do clima, solo e relevo. São as condições ideais que faz a planta ter seu melhor desenvolvimento, produzindo mais, trazendo menos risco para os produtores e evitando o abandono de áreas. E é para definir as áreas com as melhores condições que os zoneamentos agrícolas são elaborados.

O zoneamento agrícola pode levar em consideração diversos fatores, partindo dos climáticos, edáficos, sociais, econômicos. Ele é uma ferramenta importante de gestão pública e privada, sendo usado, por exemplo, no Programa de Garantia da Atividade Agropecuária (PROAGRO) e para liberação de crédito rural por agentes financeiros.

Muitos trabalhos, como de HAMADA et al. (2006), FILHO, BRAZ CALDERANO *et al.* (2004), LUMBRERAS *et al.*, (2015) utilizam a metodologia elaborada por RAMALHO-FILHO; BEEK (1995), conhecida como Sistema de Avaliação da Aptidão Agrícola das Terras. Essa metodologia baseia-se em cinco qualidades básicas (referentes à capacidade de fornecimento de nutrientes, água e oxigênio às plantas; a adequação à mecanização e a suscetibilidade aos processos erosivos), e da viabilidade de melhoramento das limitações mediante o emprego de práticas de manejo inerentes a três níveis tecnológicos, são avaliadas as possibilidades de utilização das terras com lavouras, pastagens (plantadas ou nativas) e silvicultura.

De acordo com ALKIMIM (2014) é possível utilizar um modelo de análise de multicritérios (MCDA), desenvolvido na ferramenta *Model Builder* do software ArcGIS, para identificar áreas potencias para a produção de cana-de-açúcar. No trabalho realizado por esses autores foram usados como inputs do modelo dados de solo, temperatura, precipitação, declividade, bioma, uso do solo, infraestrutura (proximidade de estradas e usinas) e dados socioeconômicos. Além disso, é importante destacar que a validação é um procedimento muito importante para garantir que um modelo seja desenvolvido corretamente.

No estudo elaborado por MENDAS; DELALI, (2012) foi desenvolvido um mapa de aptidão agrícola para o trigo *durum* na Algéria. Mesmo a aplicação desse estudo ter sido em outro país, a metodologia pode ser aplicada para outros locais. Foi utilizado um método de análise de multicritérios associado com SIG (ArcGIS), que se mostrou um poderoso sistema de suporte à decisão espacial e que oferece a oportunidade de produzir eficientemente mapas de aptidão agrícola. Os critérios usados foram a reserva de água facilmente disponível, drenagem, permeabilidade, pH, condutividade elétrica, soma de bases (V%), CTC, textura, profundidade, declividade, disponibilidade de mão-de-obra e proximidade de estradas.

Baseado nos métodos mais tradicionais, novas ferramentas vêm sendo usadas na agricultura, como é o caso da inteligência artificial (IA), que é a área da ciência da computação responsável pelo desenvolvimento de sistemas que simulam a capacidade de resolver problemas. MINSKY, (1968) fez a seguinte definição "Inteligência Artificial é a ciência de fazer as máquinas fazerem coisas que exigiriam inteligência se fossem feitas por homens." (tradução livre).

O aprendizado de máquina (*machine learning,* em inglês) "é o campo de estudo que dá aos computadores a capacidade de aprender sem ser explicitamente programado" (SAMUEL, 1959). Esse campo vem sendo muito usado no cotidiano das pessoas, através dos bancos que fazem análises de risco de crédito, nos *streams* para sugerir um filme ou série, em carros autônomos, reconhecimento de faces, diagnóstico de doenças, entre muitas outras áreas.

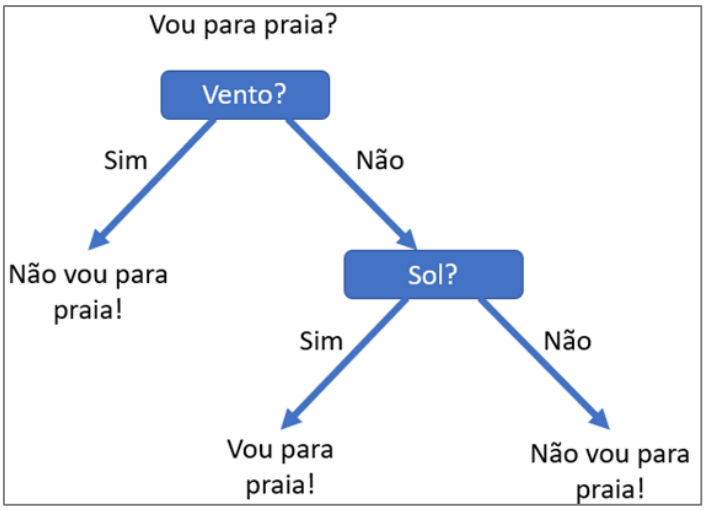
Em uma revisão bibliográfica sobre a aplicação de aprendizado de máquina na área agrícola, LIAKOS *et al.* (2018) identificou 46 artigos entre 2004 e 2018 sobre o tema, sendo 26 com agricultura, 8 com pecuária, 5 sobre manejo d'água e 7 sobre manejo do solo. Isso mostra que esse tema vem sendo estudado na agricultura e tem se mostrado relevante.

A tarefa de aprendizagem do modelo pode ser preditiva ou descritiva. Na tarefa preditiva uma previsão de resultado baseado em um conjunto de atributos é realizada, podendo ser de classificação ou regressão, no qual o resultado é uma classe (por exemplo, 0 ou 1) ou um valor numérico (por exemplo, valor da produtividade de soja), respectivamente. Já na tarefa descritiva, são buscados padrões entre os dados, eles podem ser agrupados, associados, pode ocorrer a detecção de sequências ou desvios, entre outros. Além das tarefas de aprendizagem, há também os métodos de aprendizagem, que pode ser supervisionado, não supervisionado e por reforço. Para problema de predição é utilizado a aprendizagem supervisionada, ou seja, para treinar o modelo são necessários rótulos já conhecidos para determinados conjuntos de dados. Já na aprendizagem não supervisionada, usado nos problemas descritivos, não há rótulos conhecidos, o próprio modelo cria os padrões entre os dados. Por último, o método por reforço o modelo aprende com as interações com o ambiente (causa e efeito), com sua própria experiência, esse caso é muito comum na robótica.

Existem inúmeros algoritmos de aprendizado de máquina, entre eles o *Random Forest*, tanto o estudo realizado por GUIMARÃES (2019) quanto por LORENSINI (2019) utilizaram tal algoritmo. *Random Forest* é uma técnica de classificação e regressão desenvolvida por BREIMAN (2001) que consiste em combinar várias árvores de decisão, utilizando a estratégia conhecida como *emsemble*, para obter um único resultado.

Uma árvore de decisão pode ser vista como uma representação gráfica para um determinado processo de decisão. As árvores são formadas por nós, que armazenam informação (perguntas). O nó raiz é o nó que possui maior nível hierárquico e, a partir dele, ramificam-se os nós filhos. O nó que não possui filhos é conhecido como nó folha ou terminal. A imagem abaixo exemplifica a estrutura de uma árvore. A Figura 1 apresenta um exemplo simples e prático do funcionamento de uma árvore de decisão.

Figura 1. Exemplo prático de uma árvore de decisão



Fonte: https://didatica.tech/como-funciona-o-algoritmo-arvore-de-decisao/

Para definição do atributo mais importante para iniciar a árvore e os atributos que darão sequência, o algoritmo calcula a entropia e o ganho de informação para cada atributo (Figura 2).

Figura 2. Entropia e Ganho de informação

$$Entropia(S) = \sum_{i=1}^{c} - p_i log_2 p_i$$

$$Ganho\ de\ informação\ (S, A) = Entropia(S) - \sum_{vEValues(A)} \frac{|Sv|}{|S|} Entropia(Sv)$$

Fonte: Adaptado de HURWITZ; KIRSCH (2018).

Como já abordado, o algoritmo *Random Forest* é uma melhoria do algoritmo de árvore de decisão, pois combinada a simplicidade das árvores de decisão com a aleatoriedade para melhorar a precisão. Cada árvore de decisão é construída utilizando uma amostra aleatória inicial dos dados e, a cada divisão desses dados, um subconjunto aleatório de atributos é utilizado para a escolha dos atributos mais informativos. Os principais passos do algoritmo *Random Forest* podem ser vistos na Figura 3.

Figura 3. Algoritmo básico da técnica *Random Forest*

Dado um conjunto de dados $X = x_1, x_2, \ldots, x_j$ e $Y = y_1, y_2, \ldots y_k$.

Para $b = 1, 2, 3, \ldots B$, repita:

(a) Cria uma amostra *bootstrap* (X_b, Y_b) com n exemplos de (X, Y).

(b) Ajusta uma árvore de decisão f^b para o conjunto de treinamento (X_b, Y_b), utilizando m atributos para a escolha de cada nó.

Fim de repetição.

Gera o modelo final: $\hat{f}(x) = \sum_{b=1}^{B} f^b(x)$, que calcula os votos obtidos por cada modelo f^b, resultando uma classificação final de acordo com a votação majoritária.

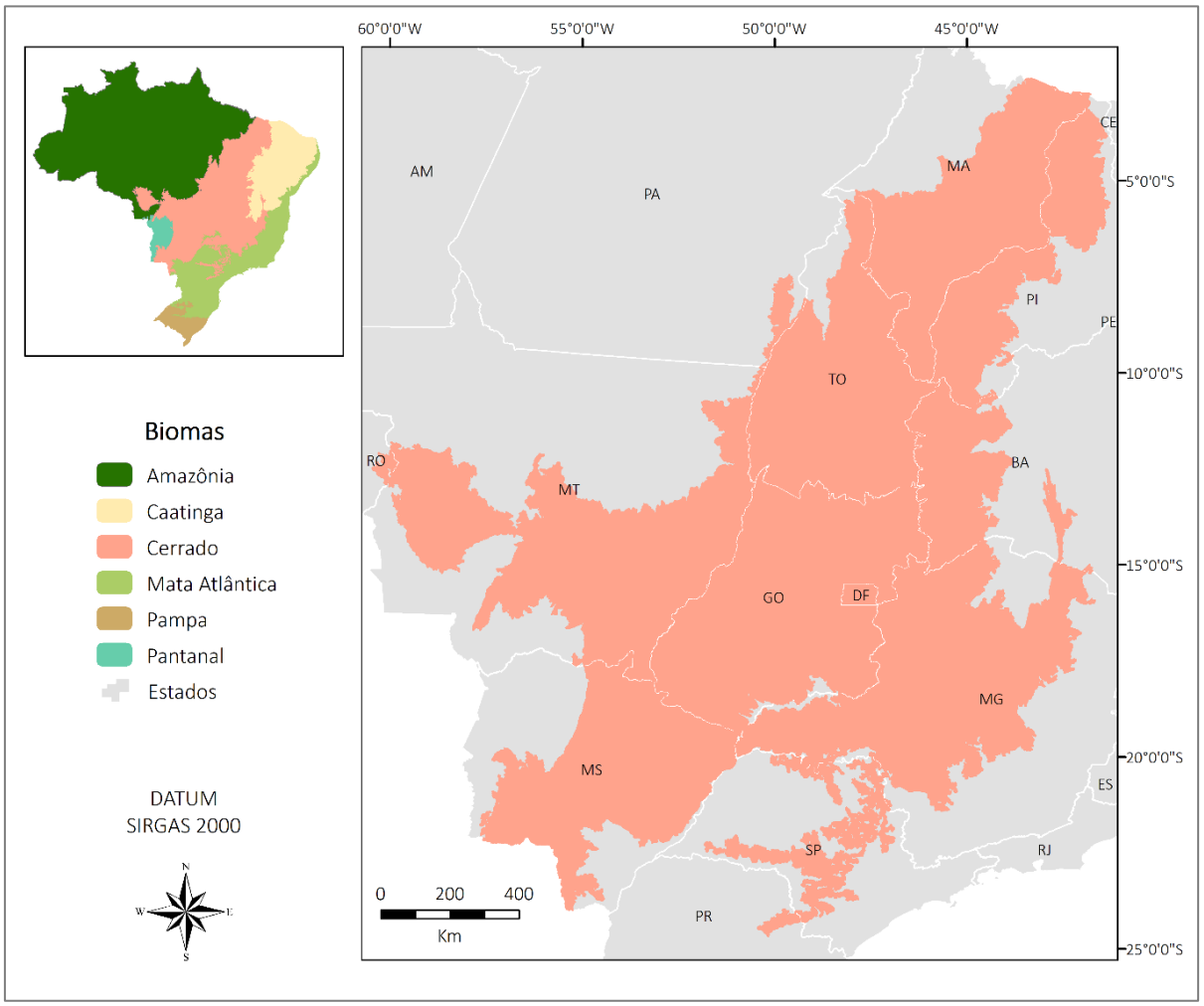
Fonte: Adaptado de VIEIRA et al. (2015)

4. MATERIAL E MÉTODO

4.1 Área de estudo

O Cerrado é o segundo maior bioma brasileiro, com uma área de 198,3 milhões de hectares, o que representa 23,3% do território do Brasil. Faz divisa com quase todos os demais biomas, exceto o Pampa (Figura 4). A precipitação varia entre 600 a 2200 milímetros anuais, já as temperaturas médias anuais variam entre 22°C a 27°C. Os solos, em sua maioria são distróficos, ácidos e com altos teores de alumínio trocável e a altitude varia de 50 m a 2000 m. A cobertura vegetal predominante é a formação savânica (IBGE, 2019).

Figura 4. Localização do bioma Cerrado



Fonte: IBGE (2019). Elaboração: Autora.

O Cerrado é formado pelos estados de Bahia, Distrito Federal, Goiás, Maranhão, Mato Grosso, Mato Grosso do Sul, Minas Gerais, Pará, Paraná, Piauí, Rondônia, São Paulo e Tocantins (Tabela 1).

Tabela 1. Área (hectares) dos estados que compõem o bioma Cerrado

Estado	Área do Estado no Cerrado (ha)	% do Estado no Cerrado
Bahia	10.339.400	18,3
Distrito Federal	576.100	100
Goiás	33.466.800	98,4
Maranhão	21.559.500	65,4
Mato Grosso	33.800.100	37,4
Mato Grosso do Sul	22.222.600	62,2
Minas Gerais	31.708.200	54,1
Pará	867.500	0,7
Paraná	312.200	0,54
Piauí	13.272.100	52,7
Rondônia	255.300	1,1
São Paulo	4.616.500	18,6
Tocantins	25.305.500	91

Fonte: Adaptado de IBGE (2019).

4.2 Base de dados

Os modelos desenvolvidos neste estudo foram construídos usando a abordagem de aprendizagem supervisionada. Isso significa que o modelo foi treinado com um conjunto de dados de entrada rotulados, no qual a saída esperada já era conhecida. A base de dados é composta pelas variáveis e rótulos, que foram chamados de atributos previsores e classes, respectivamente.

4.2.1 Atributos previsores

Os atributos previsores são os dados de entrada para o algoritmo, o qual aprende a relação entre eles para resultar em um determinado rótulo. Nesse estudo foram usados atributos que possuem alguma relação com potencial agrícola, entre eles, atributos edafoclimáticos, de infraestruturas e socioeconômicos (Tabela 2). A escolha dos atributos previsores foi baseada no estudo desenvolvido por HARFUCH et al. (2021).

Tabela 2. Atributos previsores usados nos modelos

	Atributos	Fonte	Ano	Formato	Referência
Edafoclimático	Solos	Soil Grids	2020	Raster	https://soilgrids.org/
	Temperatura média mensal	WorldClim	1970 - 2000	Raster	https://www.worldclim.org/data/worldclim21.html
	Temperatura máxima mensal	WorldClim	1970 - 2000	Raster	
	Temperatura mínima mensal	WorldClim	1970 - 2000	Raster	
	Precipitação mensal	WorldClim	1970 - 2000	Raster	
	Altitude	EMBRAPA	2005	Raster	http://www.relevobr.cnpm.embrapa.br/
	Declividade	TOPODATA	2011	Raster	http://www.webmapit.com.br/inpe/topodata/
Infraestrutura	Ferrovias	Ministério da Infraestrutura	2020	Shapefile	http://www.infraestrutura.gov.br/component/content/article/63-bit/5124-bitpublic.html#maprodo
	Rodovias	Ministério da Infraestrutura	2020	Shapefile	
	Hidrovias	Ministério da Infraestrutura	2020	Shapefile	
	Portos	Ministério da Infraestrutura	2020	Shapefile	
	Frigoríficos	Lapig	2017	Shapefile	https://www.lapig.iesa.ufg.br/lapig/index.php/produtos/dados-geograficos
	Armazéns	Conab	2019	Tabela	https://www.conab.gov.br/armazenagem
	Siderúrgicas	Instituto Aço Brasil	2018	Tabela	https://institutoacobrasil.net.br/site/parque-siderurgico/
	Processamento de tora	CERFLOR, ABIMCI, IBA	2018	Tabela	-
Socioeconômico	Ocupação no setor do Agronegócio	Censo Demográfico	2010	Tabela	http://www.atlasbrasil.org.br/perfil
	População total	Censo Demográfico	2010	Tabela	
	População urbana	Censo Demográfico	2010	Tabela	
	População rural	Censo Demográfico	2010	Tabela	
	Índice de Desenvolvimento Humano (IDH)	Censo Demográfico	2010	Tabela	
	Renda per capita	Censo Demográfico	2010	Tabela	
	Produção de soja anual	PAM - IBGE	2010 - 2019	Tabela	http://www.sidra.ibge.gov.br
	Efetivo de bovinos anual	PPM – IBGE	2010 - 2019	Tabela	
	Produção de produtos madeireiros anual	PEVS – IBGE	2010 - 2019	Tabela	

Os dados de clima, relevo e solo são os chamados atributos edafoclimáticos. Foram utilizados os dados mensais de precipitação, radiação solar, temperatura média, temperatura máxima e temperatura mínima da base do WorldClim (FICK; HIJMANS, 2017). Juntamente foram utilizados os dados de altitude oriundos da Embrapa (MIRANDA, 2005), os dados de declividade em porcentagem do TopoData (VALERIANO, 2008) e os dados de solo do *SoilGrids* (HENGL *et al.*, 2017), os últimos foram obtidos através da plataforma do *Google Earth Engine*.

Os atributos de infraestrutura usados no estudo são compostos pela localização das rodovias federais e estaduais, ferrovias, hidrovias e portos, ambos do Ministério da Infraestrutura (2020). Também foram utilizados os dados de frigoríficos (LAPIG, 2019), silos e armazéns (CONAB, 2019), siderúrgicas e locais de processamento de tora (IBÁ, CERFLOR, ABIMCI, 2018). O intuito foi incluir dados relacionados com as cadeias agropecuárias analisadas.

Por fim, foram usados os dados do Censo Demográfico (IBGE, 2011), da Produção Agrícola Municipal (PAM/IBGE, 2018), da Produção da Pecuária Municipal (PPM/IBGE, 2018) e da Produção da Extração Vegetal e da Silvicultura (PEVS/IBGE, 2019), que foram chamados de atributos socioeconômicos.

Após a aquisição dos dados, todos foram padronizados utilizando o ArcMap, que é um aplicativo integrado do sistema de informação geográfica ArcGIS. A padronização compreendeu em ter todos os dados em formato matricial (*raster*), com resolução espacial de 1.000 metros e no *datum* SIRGAS 2000.

Os atributos edafoclimáticos já foram obtidos em formato matricial (*raster*), eles foram reamostrados para uma resolução espacial de 1.000 metros, utilizando a ferramenta *Resample*, e transformado para o *datum* SIRGAS 2000, utilizando a ferramenta *Project Raster*.

Os dados tabulares de infraestrutura foram espacializados a partir de coordenadas geográficas ou endereço, obtendo um arquivo vetorial (*shapefile*) de pontos. Com todos os dados de infraestrutura no formato vetorial, foi calculada a distância euclidiana utilizando a ferramenta *Euclidean Distance*, obtendo dados matriciais que foram padronizados. E, por último, os dados socioeconômicos foram espacializados por municípios, transformados para formato matricial utilizando a ferramenta *Feature to Raster* e padronizados.

4.2.2 Classe

A classe é o dado utilizado como rótulo para treinar o modelo, é a informação conhecida que será passada para o algoritmo de aprendizagem supervisionada. Este dado pode ser uma informação real de campo – no contexto de potencial agrícola - ou uma informação gerada por outros estudos. Nesta pesquisa a segunda opção se aplicou, foi usado como base o estudo de HARFUCH *et al.* (2021).

HARFUCH *et al.* (2021) apresentou as áreas potenciais para expansão das cadeias da soja, pecuária de corte, pecuária de leite e floresta plantada sobre pastagens degradadas. De acordo com o estudo, as áreas potenciais para soja são aquelas que possuem aptidão agrícola, estão num raio de 20 km das áreas já existente de soja e dos silos, e possuem uma área contínua de, no mínimo, 100 ha. Já as áreas com potencial para pecuária de corte e de leite são aquelas que estão num raio de 150 km dos frigoríficos e 100 km dos laticínios, respectivamente. Por fim, as áreas potenciais para floresta plantada são aquelas num raio de 20 km das áreas já existentes de floresta plantada (HARFUCH *et al.*, 2021). Há outros estudos que apresentam a oportunidade de expansão para soja, que utilizam os mesmos critérios citados acima (ROMEIRO *et al.*, 2018)(COSTA *et al.*, 2021).

As premissas descritas acima foram adotadas na elaboração dos dados usados como classe para treinar o modelo. Foram gerados quatros diferentes *rasters*: i. Potencial de expansão da soja; ii. Potencial de intensificação da pecuária; iii. Potencial de expansão da floresta plantada e iv. Compilado das três cadeias agropecuárias.

O dado de potencial de expansão da soja foi desenvolvido usando a base de aptidão agrícola na pastagem e na vegetação nativa (RUDORFF; RISSO, 2015), de silos (CONAB, 2019) e de área de soja em 2019 ("MapBiomas - Coleção 5", 2020). Foram gerados raios de 20 km dos silos e das áreas de soja utilizando a ferramenta *Buffer* e com a ferramenta *Clip* foram recortadas as áreas com alta e média aptidão, e sem restrição de declividade dentro desses raios. As áreas resultantes destes processos foram somadas as áreas de soja em 2019, e classificadas como "potencial para expansão da soja". As demais áreas foram classificadas como "sem potencial".

Para o dado de potencial de intensificação da pecuária, foram selecionadas como potencial as áreas de pastagem em 2019 (LAPIG, 2019) que estão inseridos no raio de 150 km dos frigoríficos (LAPIG, 2019). Estas áreas foram classificadas como "potencial de intensificação da pecuária" e as demais como "sem potencial".

E, por último, foram classificadas como "potencial para expansão da floresta plantada" toda a área inserida no raio de 20 km das áreas de floresta plantada estabelecidas em 2019 ("MapBiomas - Coleção 5", 2020). O restante foi classificado como "sem potencial".

Ao juntar todos os *rasters* elaborados, foi identificado a concorrências de áreas entre mais de uma cadeia (soja, pecuária e/ou floresta plantada). Por isso, um novo dado foi criado com cinco classes: soja, pecuária, floresta plantada e sobreposição.

Com os atributos previsores e as classes em formato matricial, uma malha de pontos com distância de 1 quilômetro entre eles foi gerada, utilizando a função *Raster to point*. Em seguida, foram extraídas as informações de cada camada utilizando a ferramenta *Extract Multi Values to Points* e adicionado as informações de latitude e longitude com a ferramenta *Add XY Coordinates*. Por fim, essa malha de pontos foi exportada para o formato *csv* (valores separados por vírgula).

4.3 Modelagem

Toda a parte de modelagem foi realizada em linguagem *Python*, utilizando o *Spyder* do *Anaconda Navigator*, que é um *open source* desenvolvido em linguagem *Python* para executar *Python*.

Foram elaborados quatro diferentes modelos, o primeiro foi um modelo para identificar o potencial de expansão da soja, para isso foram usados todos os atributos como dados de entrada e, como classe foi usado o dado de potencial de expansão da soja apresentado acima, que indica se uma determinada localização tem potencial ou não. Este último é o dado com o qual o algoritmo vai aprender a combinação de atributos que resulta em áreas com potencial e sem potencial.

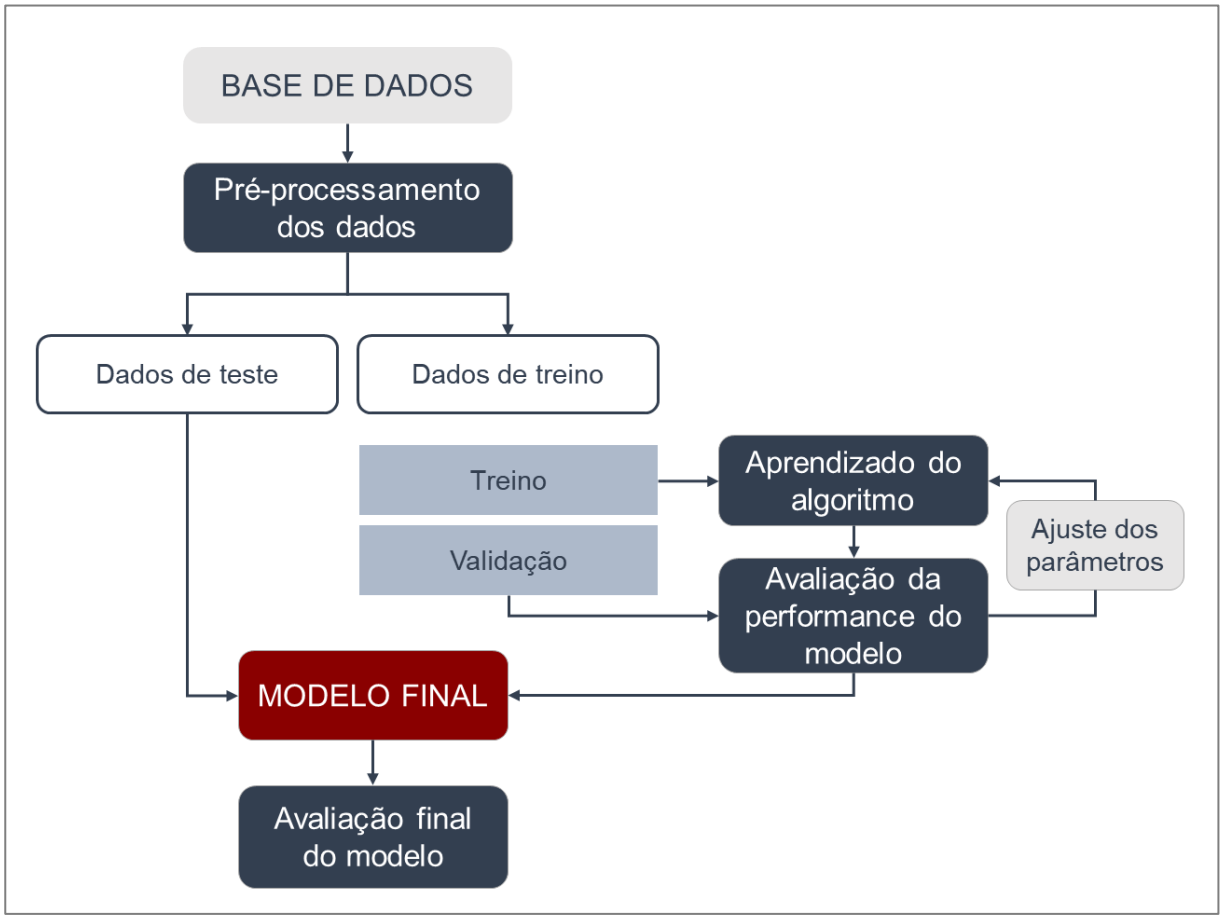
A mesma lógica foi utilizada para desenvolver o segundo e terceiro modelos, que tiveram como objetivos indicar o potencial de intensificação da pecuária e expansão da floresta plantada, respectivamente.

O último, denominado modelo final, foi desenvolvido com o intuito de ter um único modelo que fosse capaz de classificar o potencial para as três cadeias (soja, pecuária e floresta plantada) sem que haja sobreposição entre elas, além de indicar as áreas sem potencial. Para isso, foram usados todos os atributos e, como classe, o dado que contém as classes de soja, pecuária, floresta plantada, sem potencial e com sobreposição, citado anteriormente.

4.3.1 Fluxo de trabalho

O fluxo de trabalho da modelagem é apresentado na Figura 5. Com a base de dados padronizada e organizada, a etapa de modelagem pode ser iniciada. O primeiro passo foi realizar o pré-processamento dos dados, em sequência estes dados foram divididos entre dados de treino e teste. O algoritmo de aprendizado de máquina foi treinado, foram feitos ajustes nos parâmetros para melhorar seu desempenho, o modelo foi aplicado aos dados de teste e, por fim, teve sua performance final avaliada.

Figura 5. Fluxograma de trabalho da modelagem



Fonte: Resultado do estudo. Elaboração: autora.

4.3.2 Pré-processamento dos dados

O pré-processamento consiste em tratar os dados para as fases seguintes e, também pode ser útil para o melhor conhecimento da base de dados. Nesta etapa, várias funções podem ser aplicadas, como correção de dados faltantes e de outliers, seleção de atributos mais relevantes, normalização dos dados, análise de correlação, entre outros.

Neste estudo, a primeira etapa de pré-processamento foi corrigir os dados faltantes, para isso foi calculado a média de cada variável e os dados faltantes foram substituídos por esse valor. Em seguida, as linhas da classe sobreposição foram excluídas, pois não era de interesse que o algoritmo aprendesse a identificar esta classe.

Após o pré-processamento, a base de dados foi dividida entre dados de treino e de teste, usando a proporção de 80% e 20%, respectivamente (SHALEV-SHWARTZ; BEN-DAVID, 2013). Os dados de treino foram utilizados tanto para o treinamento do modelo, quanto para a validação dele. Já os dados de teste foram utilizados para a verificação da performance final.

4.3.3 Modelo

Os algoritmos de aprendizado de máquina testados e as ferramentas utilizadas nesta parte do estudo são da biblioteca do *Scikit-learn*, que é um "módulo Python que integra uma vasta gama de algoritmos de aprendizagem de máquinas de última geração para problemas supervisionados e não supervisionados a média escala" (PEDREGOSA *et al.*, 2011).

Foram desenvolvidos 4 modelos de classificação usando o algoritmo *Random Forest*:

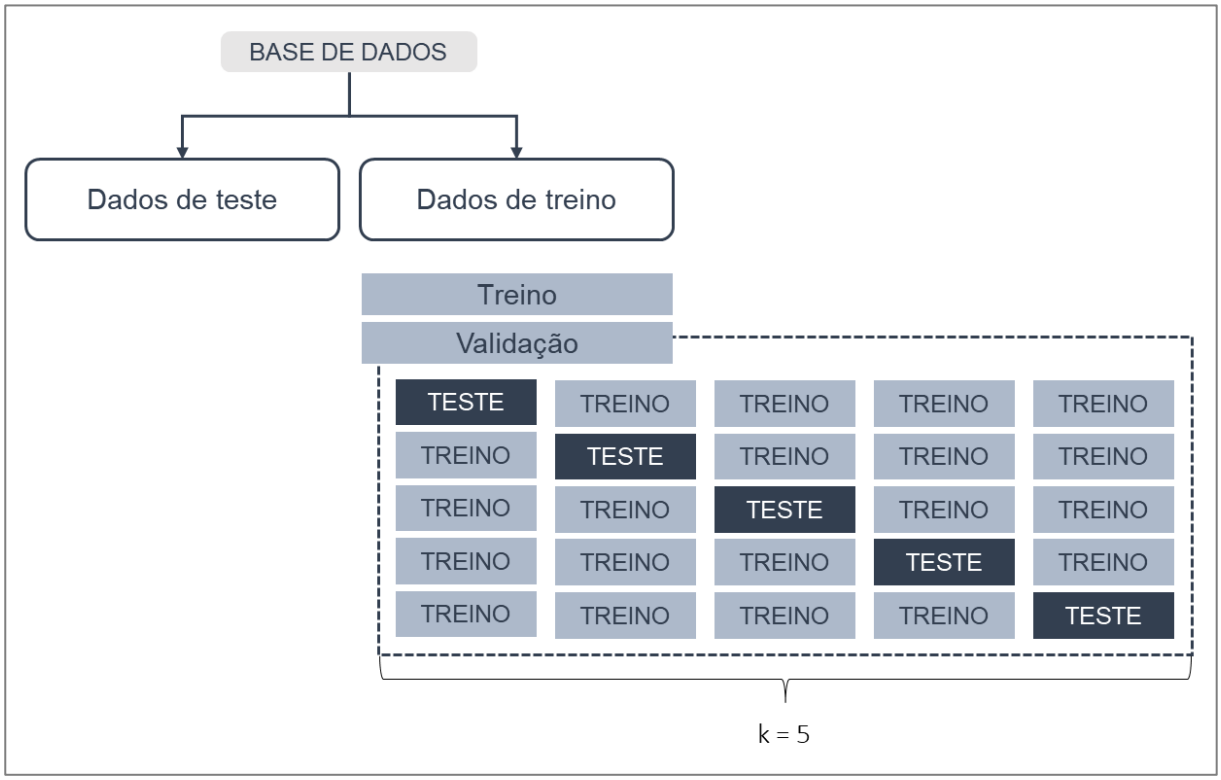
- Potencial de expansão da soja, com 2 classes – com potencial e sem potencial.
- Potencial de intensificação da pecuária, com 2 classes – com potencial e sem potencial.
- Potencial de expansão da floresta plantada, com 2 classes – com potencial e sem potencial.
- Potencial para as 3 atividades agrícolas acima, com 4 classes – expansão da soja, intensificação da pecuária, expansão da floresta plantada e sem potencial.

4.3.4 Validação cruzada e ajuste dos parâmetros

A validação cruzada pode ter dois objetivos principais: avaliar a performance de um modelo e sua capacidade de generalização, comparar a performance entre dois ou mais modelos e para ajuste dos parâmetros (BAK; LENNOX, 2005; WATANABE, 2009).

Neste estudo, a validação cruzada teve o objeto de avaliar a performance do modelo e auxiliar no ajuste dos parâmetros do algoritmo. Dessa forma, foi aplicado o método *k-fold*, que de acordo com SHALEV-SHWARTZ; BEN-DAVID (2014) consiste em dividir os dados em k partes, sendo k-1 os dados de teste e os demais dados de treino. O algoritmo é executado *k* vezes e se obtém um valor de acurácia para cada divisão, quanto mais próximo esses valores, melhor a capacidade de generalização do modelo e menor o *overfitting*. A Figura 6 apresenta um exemplo de validação de dados por *k-fold*, sendo o k = 5.

Figura 6. Validação cruzada k-fold



Fonte: Adaptado de SANTANA (2020). Elaboração: Autora.

O algoritmo *Random Forest* tem 18 parâmetros que podem ser alterados para melhorar sua performance com determinado conjunto de dados. De acordo com PEDREGOSA et al. (2011), os principais parâmetros do algoritmo são:

- *n_estimator* que corresponde ao número de árvores, sendo que quanto maior, melhor o modelo e, também, maior o tempo de processamento. Há um ponto que o aumento no número de árvores não resulta numa melhora expressiva da acurácia, dessa forma, é interessante buscar o número ideal que combine boa acurácia e menor tempo de execução.
- *Max_features* que corresponde ao número atributos considerados na divisão de um nó. Para problemas de regressão é recomendado o uso do *max_features* = *None*, no qual são considerados todos os atributos. Já para problemas de classificação, é recomendado a utilização do *max_features* = *sqrt*, no qual é utilizado um número de atributos igual a raiz quadrada do número total de atributos.

Diversos valores podem ser testados para cada parâmetro, isso pode ser uma atividade trabalhosa e que demanda tempo. Para facilitar esta etapa foi utilizado a função *GridSearch*, a qual permite estabelecer diferentes valores para os parâmetros, que são testados exaustivamente entre si, de forma que seja encontrada a melhor combinação de valores, resultando numa melhor performance do modelo.

A combinação dos melhores valores, pode resultar numa melhor acurácia, porém ao mesmo tempo é necessário avaliar a capacidade de generalização desse modelo. Por isso, após a escolha dos valores, o modelo é validado pelo método *k-fold* (descrito acima) e, em seguida, os valores dos parâmetros são ajustados novamente, entrando em um looping até que seja feito o melhor ajuste.

4.3.5 Avaliação final do modelo

O modelo gerado foi aplicado nos dados de teste, a partir daí foi possível fazer a avaliação final da performance do modelo. Para tal, foi utilizado a função *confusion_matrix* para geral a Matriz de Confusão, que apresenta os valores preditos comparados com os valores reais (Figura 7).

Quando o modelo é binário, ou seja, com duas classes, a matriz de confusão é composta por quatro valores: verdadeiros positivos (VP) que indica o número de previsões corretas da classe positivo; falsos positivos (FP) indica a quantidade de previsões positivas, mas que no valor real são negativas; falsos negativos (FN) são os valores que o modelo prevê que é da classe negativa, mas é da classe positiva; e verdadeiros negativos (VN) que indica os acertos da classe negativa.

Figura 7. Matriz de confusão

		Valores Reais	
		Positivo	Negativo
Previsões	Positivo	VP	FP
	Negativo	FN	VN

Fonte: Adaptado de DAVIS et al. (2016). Elaboração: Autora.

A partir destes valores é possível calcular outras métricas que auxiliam na avaliação da performance do modelo, entre eles:

Acurácia é a indicação geral de quanto o modelo acertou, ela é dada pela divisão da somatória dos números verdadeiros pelo total (1).

$$Acurácia = \frac{VP + VN}{VP + FP + VN + FN} \quad (1)$$

Precisão é a relação entre o número de acerto de uma classe pelo número total de dados dela. A precisão pode ser calculada tanto para a classe positiva (2) quanto a negativa (3).

$$Precisão\ (positivo) = \frac{VP}{VP + FP} \quad (2)$$

$$Precisão\ (negativo) = \frac{VN}{VN + FN} \quad (3)$$

Sensibilidade ou taxa de verdadeiro positivo (TVP) indica o percentual que foi predito positivo sobre o que realmente era positivo (Equação (4).

$$Sensibilidade = \frac{VP}{VP + FN} \quad (4)$$

Especificidade ou taxa de verdadeiro negativo (TFN) indica o percentual que foi predito negativo sobre o que realmente é negativo (Equação (5).

$$Especificidade = \frac{VN}{VN + FP} \tag{5}$$

Taxa de falso positivo (TFP) indica o percentual de predições positivas que na realidade são negativas (Equação 6).

$$TFP = \frac{FP}{VN + FP} \tag{6}$$

Taxa de falso negativo (TFN) indica o percentual de predições negativas que na realidade são positivas (Equação 7).

$$FN = \frac{FN}{FN + VP} \tag{7}$$

5. RESULTADOS E DISCUSSÃO

5.1 Parâmetros dos modelos

Os primeiros testes para o treinamento do modelo foram realizados com os parâmetros Default. Posteriormente, alguns parâmetros foram sendo ajustados, levando em consideração os mais importantes, a fim de aumentar a acurácia do modelo. Dessa forma, ficaram estabelecidos os valores apresentados na Tabela 2.

Tabela 2. Valores dos parâmetros para cada modelos

Parâmetros	Soja	Pecuária	Floresta	Final
N_estimators	10	50	10	10
Criterion	Default	Default	Default	Default
Max_depth	15	20	15	20
Min_sample_split	Default	Default	Default	Default
Min_sample_leaf	Default	Default	Default	10
Min_weight_fraction_leaf	Default	Default	Default	Default
Max_features	Default	Default	Default	Default
Max_leaf_nodes	Default	Default	Default	Default
Min_impurity_decrease	Default	Default	Default	Default
Bootstrap	Default	Default	Default	Default
Oob_score	Default	Default	Default	Default
N_jobs	Default	Default	Default	Default
Random_state	0	0	0	0
Verbose	Default	Default	Default	Default
Warm_state	Default	Default	Default	Default
Class_weigth	Balanced	Balanced	Balanced	Balanced
Ccp_alpha	Default	Default	Default	Default
Max_samples	Default	Default	Default	Default

Fonte: Resultado do estudo. Elaboração: Autora.

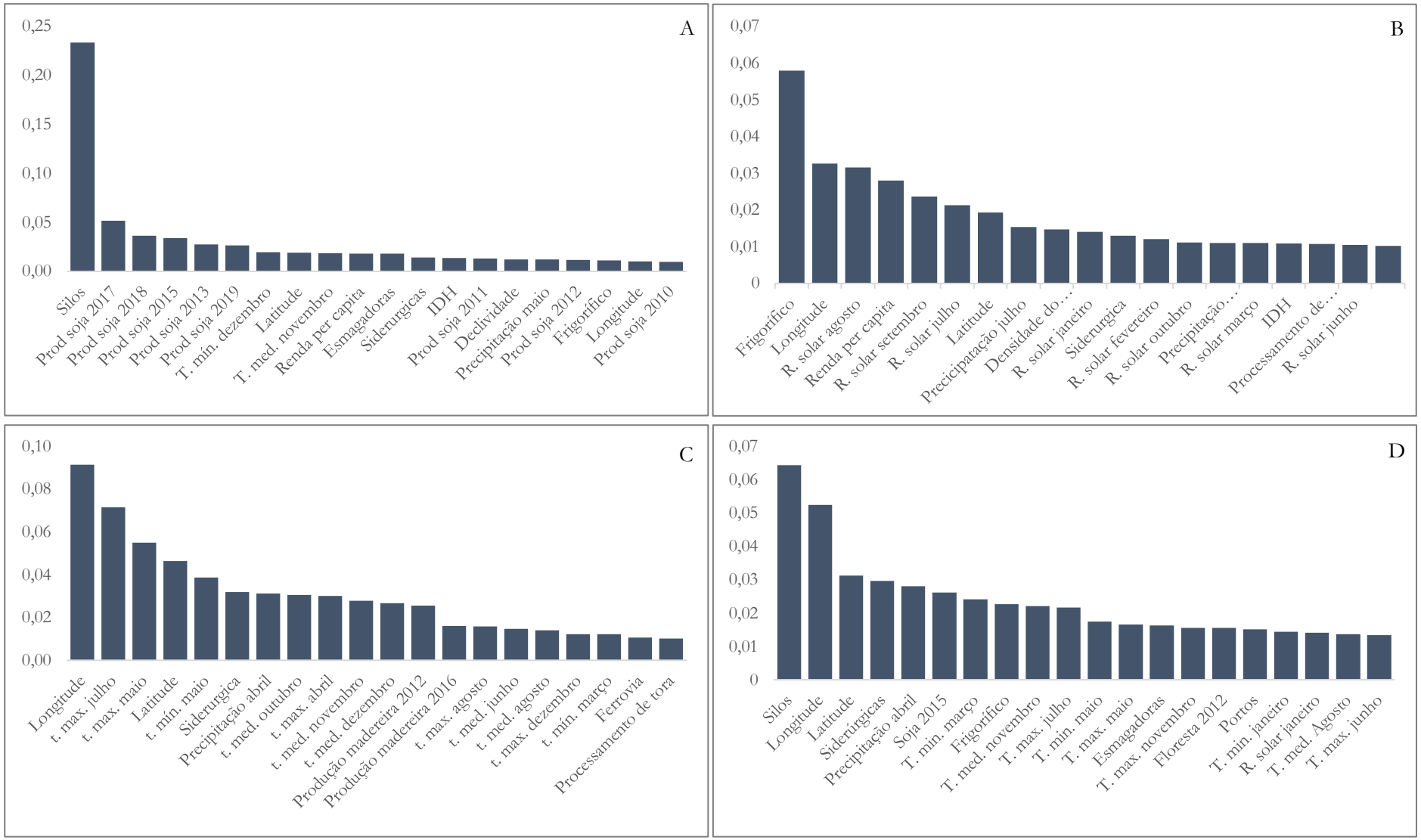
5.2 Importância das variáveis

A importância das variáveis é uma das possíveis saídas do algoritmo *Random Forest*, é dado um valor para cada variável, sendo que quanto maior, mais importante a variável é. A soma dos valores de todas as variáveis é 1.

Na Figura 8 são apresentados 4 gráficos com as 20 variáveis mais importante de cada modelo. Nos modelos de potencial para expansão da soja e intensificação da pecuária, a variável mais importante foi aquela relacionada à indústria do setor, silos e frigoríficos, respectivamente. Isso se deve ao fato de que um dos critérios para indicar o potencial dessas cadeias foi a proximidade com tais indústrias.

Para o modelo de potencial de expansão da floresta plantada, a indústria não foi a principal variável, mas está listada entre as 20 mais importantes. Já no modelo final, que engloba as três cadeias agropecuária, as variáveis de destaque foram silos, latitude e longitude.

Figura 8. Importância das variáveis de cada modelo (a) potencial de expansão da soja (b) potencial de intensificação da pecuária, (c) potencial de expansão da floresta plantada (d) potencial das cadeias agropecuárias



5.3 Performance dos modelos

Neste item serão apresentados a avaliação da performance do modelo, o que inclui dados de acurácia, sensibilidade, número de verdadeiros positivos, verdadeiros negativos, entre outros. Incluso são apresentados dados espacializados e valores de áreas, que ajudam no entendimento das estatísticas.

Para facilitar a exposição dos dados, foram usados dois termos: dados reais e dados preditos. Os dados reais são aqueles que foram gerados usando as premissas do estudo de HARFUCH *et al.* (2021), descrito no item 4.2.2 deste estudo. Já os dados preditos são aqueles gerados pelos modelos de aprendizado de máquina desenvolvidos.

5.3.1 Soja

A avaliação final da performance do modelo de expansão de soja mostrou uma acurácia geral de 86%. Analisando detalhadamente cada valor da matriz, é possível observar a sensibilidade do modelo foi de 93%, ou seja, de todos os pontos classificados como "com potencial", 93% estão corretos. Por outro lado, a precisão dos valores verdadeiros chegou a 63%, isso significa que o modelo acabou classificando mais áreas "com potencial" que na verdade são "sem potencial". (Tabela 3).

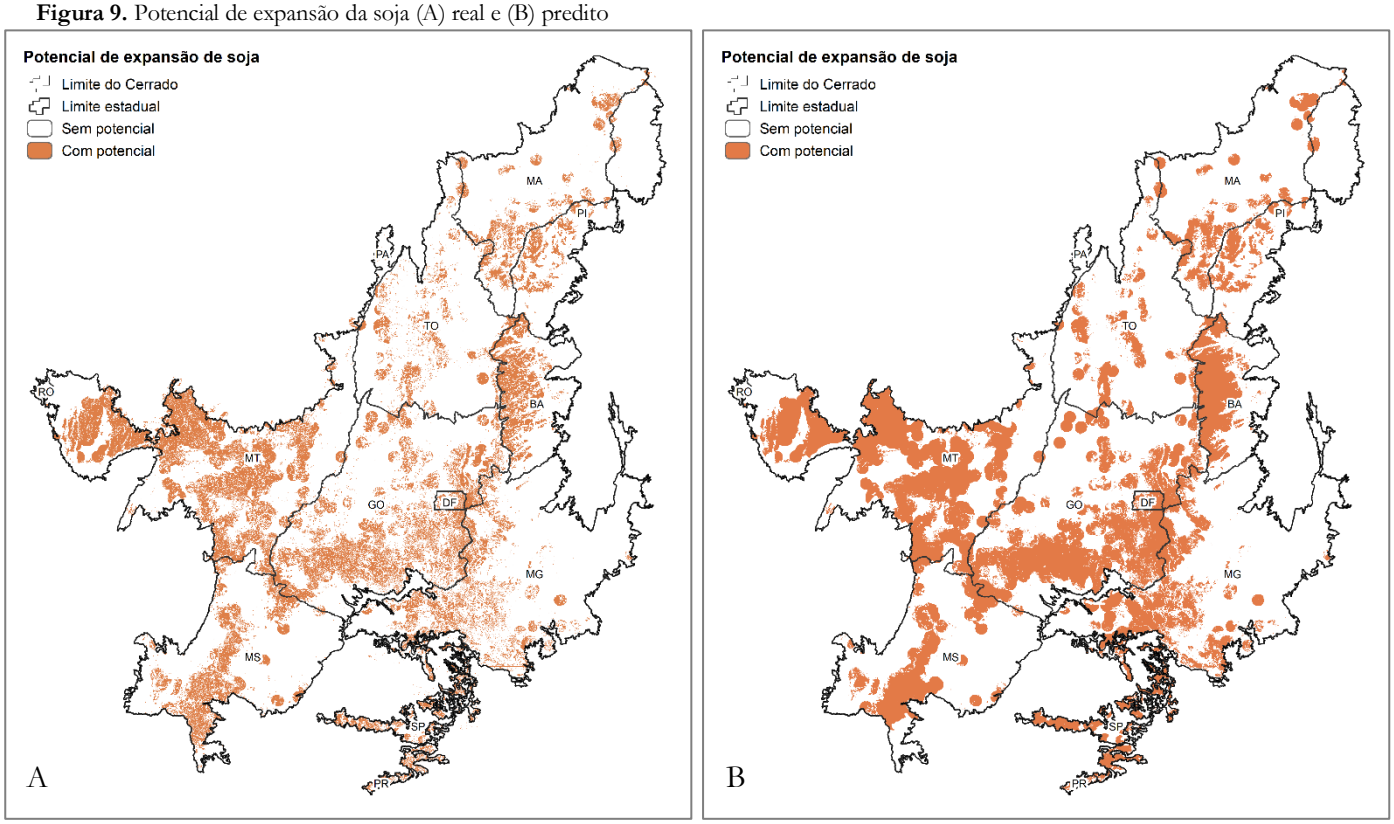
Tabela 3. Matriz de confusão do modelo de potencial de expansão da soja

		Valor real		
		Com potencial	**Sem potencial**	
Valor predito	**Com potencial**	Verdadeiro Positivo	Falso Positivo	Precisão positivo
		93.471	54.665	63%
	Sem potencial	Falso Negativo	Verdadeiro Negativo	Precisão negativo
		7.330	289.705	98%
		Sensibilidade	Especificidade	Acurácia
		93%	84%	86%

Fonte: Resultado do estudo. Elaboração: Autora

O modelo foi aplicado a todos os pontos do bioma Cerrado que, em seguida, foram espacializados. Dessa forma, foi possível comparar espacialmente os dados reais com os dados preditos (Figura 9). A partir da análise dos mapas, é possível observar que o modelo é capaz de indicar grandes extensões de áreas com potencial para soja, nas mesmas regiões que o dado real indica. Entretanto, ele não é capaz de apresentar as informações com um maior nível de detalhe, como o mapa de dado real. A generalização das áreas explica a precisão de verdadeiros positivos ser de 63%. Ao comparar os mapas de dado real com o dado predito, a acurácia foi de 87%, bem próxima da acurácia do modelo.

Figura 9. Potencial de expansão da soja (A) real e (B) predito



Fonte: Resultados do estudo. Elaboração: Autora.

Em questão de área, o modelo classificou uma área com potencial maior que a área real, isso se deve a generalização que o modelo faz, como visto na figura anterior. A área real com potencial no Cerrado é de 42,3 milhões de hectares, enquanto a área predita soma 60,7 milhões. As maiores diferenças ocorrem nos estados de Goiás e Mato Grosso, como apresentado na Tabela 4.

Tabela 4. Área (milhões de hectares) sem e com potencial para expansão da soja - dados reais e preditos

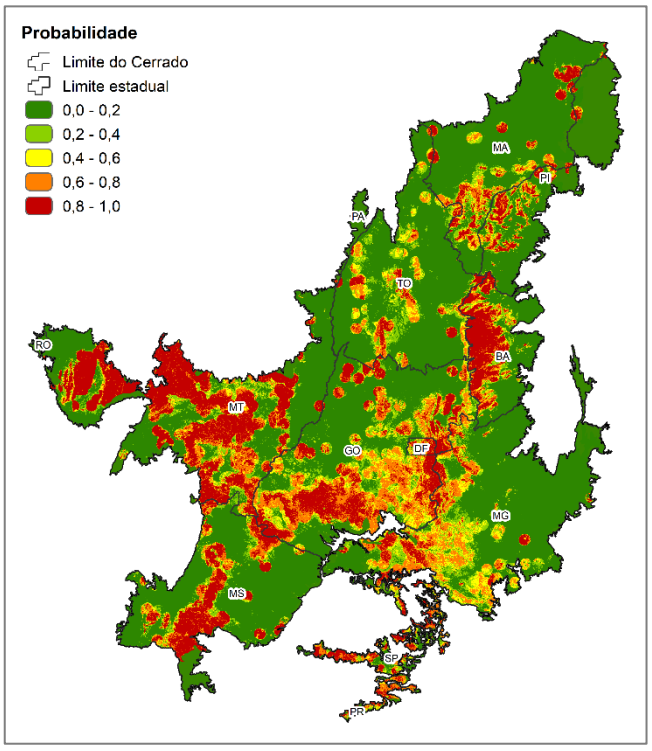
Estado	Sem potencial		Com potencial	
	Real	Predito	Real	Predito
BA	7,1	5,8	3,2	4,5
DF	0,3	0,2	0,2	0,4
GO	24,3	18,7	9,1	14,8
MA	18,9	18,0	2,3	3,3
MG	27,3	24,4	4,3	7,2
MS	17,6	16,4	4,6	5,8
MT	21,1	17,4	12,5	16,2
PA	0,8	0,8	0,0	0,0
PI	11,5	11,0	1,5	2,0
PR	0,2	0,2	0,1	0,2
RO	0,1	0,1	0,0	0,0
SP	2,8	1,8	1,6	2,6
TO	22,5	21,5	2,8	3,8
Total	**154,6**	**136,2**	**42,3**	**60,7**

Fonte: Resultados do estudo. Elaboração: Autora.

Ademais, há outro produto gerado pelo modelo, que é a predição de probabilidade. Isso significa que ao invés do modelo classificar uma área com potencial ou sem potencial, ele indica a probabilidade de ela ter potencial. É bom destacar que a aplicação do modelo aos dados para gerar um resultado binário (com potencial e sem potencial) tem como padrão o corte de 50% de probabilidade. Este tipo de dado permite uma análise mais crítica dos resultados e, também, possibilita fazer diferentes cortes e selecionar as melhores áreas.

A Figura 10 apresenta o mapa de probabilidade de potencial para expansão da soja com 5 classes: 0 – 20%, 20 – 40%, 40 – 60%, 60 – 80% e 80 – 100%. Aproximadamente, 35 milhões de hectares tem mais de 80% de probabilidade de potencial para expansão da soja. Isso pode indicar que se fossem testados diferentes cortes de probabilidade, o modelo poderia ter uma melhor acurácia e menor número de falsos positivos.

Figura 10. Mapa de probabilidade de potencial para expansão da soja



Fonte: Resultado do estudo. Elaboração: Autora.

5.3.2 Pecuária

O modelo de potencial de intensificação da pecuária apresentou uma acurácia de 79%, sendo que a precisão de verdadeiros positivos foi de 60% e sensibilidade de 85% (Tabela 5). Em comparação com o modelo de potencial de expansão de soja, ele teve uma menor performance, isso pode ter ocorrido pelas áreas de pastagens serem mais dispersas pelo Cerrado. Além disso, a base que gerou os dados de treino para a pecuária só tem como parâmetro a proximidade que as áreas de pastagens estão dos frigoríficos. Diferentemente dos dados de soja, que foram usados mais parâmetros, como proximidade de silos e de áreas de soja, e aptidão agrícola.

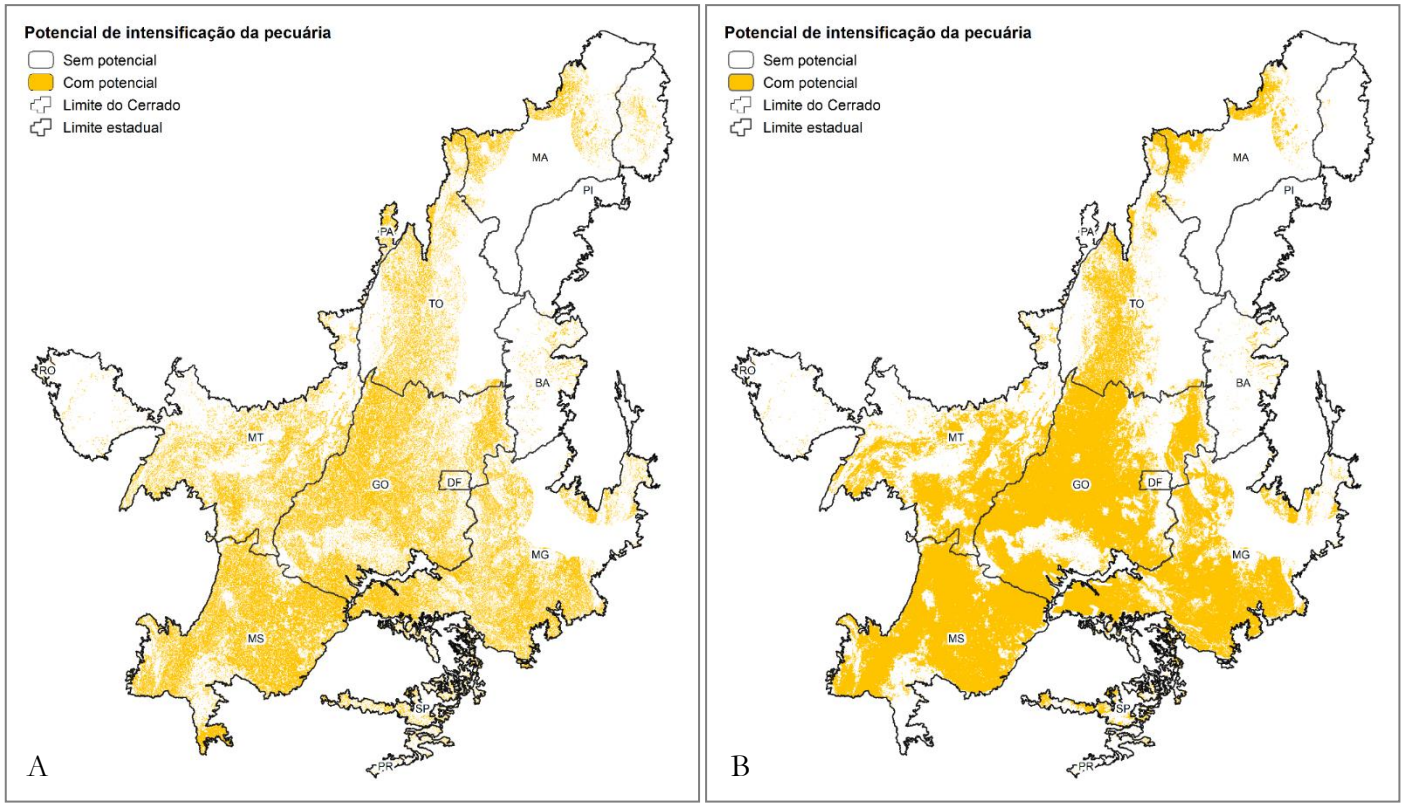
Tabela 5. Matriz de confusão do modelo de intensificação da pecuária

		Valor real		
		Com potencial	**Sem potencial**	
Valor predito	**Com potencial**	Verdadeiro Positivo	Falso Positivo	Precisão positivo
		109.140	73.703	60%
	Sem potencial	Falso Negativo	Verdadeiro Negativo	Precisão negativo
		19.930	242.398	92%
		Sensibilidade	Especificidade	Acurácia
		85%	77%	79%

Fonte: Resultados do estudo. Elaboração: Autora.

O modelo foi aplicado em todos os pontos do Cerrado, a fim de comparar os valores preditos com os reais. É possível identificar grandes regiões para intensificação da pecuária, mas de forma genérica e, não detalhada como nos dados reais (Figura 11). Esse é o mesmo comportamento visto no modelo de potencial de expansão de soja.

Figura 11. Potencial de intensificação da pecuária (A) real e (B) predito



Fonte: Resultados do estudo. Elaboração: Autora.

O dado da área com potencial predita foi 35% (19,6 milhões de hectares) maior que a área real, igual foi observado no modelo de potencial de expansão da soja. Isso demonstra que o modelo de potencial de intensificação da pecuária também classifica grande áreas com potencial, não tendo os detalhes que os dados reais apresentam. As

maiores diferenças de áreas foram observadas nos estados de Goiás, Minas Gerais, Mato Grosso e Mato Grosso do Sul (Tabela 6).

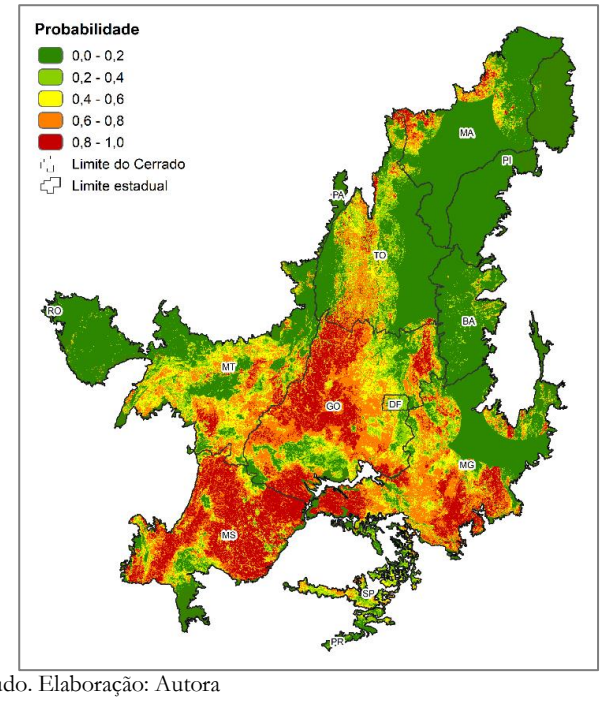
Tabela 6 - Área (milhões de hectares) sem e com potencial para intensificação da pecuária - dados reais e preditos

Estado	Sem potencial		Com potencial	
	Real	Predito	Real	Predito
BA	9,8	10,0	0,5	0,3
DF	0,4	0,4	0,1	0,2
GO	18,0	10,4	15,4	23,1
MA	19,6	19,2	2,1	2,5
MG	20,8	16,2	10,9	15,5
MS	9,6	5,3	12,7	16,9
MT	26,4	23,6	7,4	10,2
PA	0,5	0,9	0,4	0,0
PI	13,0	13,3	0,3	0,0
PR	0,3	0,3	0,0	0,0
RO	0,2	0,3	0,0	0,0
SP	3,5	3,7	1,1	0,9
TO	20,3	19,0	5,0	6,3
Total	**142,3**	**122,6**	**56,1**	**75,7**

Fonte: Resultado estudo. Elaboração: Autora.

A Figura 12 mostra o mapa de probabilidade de potencial de intensificação da pecuária. Cerca de 29 milhões de hectares tem mais de 80% de probabilidade. Mais uma vez, caso fosse adotado outro ponte de corte, diferente do 50% de probabilidade (default do modelo), poderia ser possível obter um modelo mais conservador na classificação das áreas e mais próximo das áreas reais.

Figura 12. Mapa de probabilidade do potencial de intensificação da pecuária



Fonte: Resultado do estudo. Elaboração: Autora

5.3.3 Floresta

O modelo de potencial de expansão da floresta plantada foi o que obteve a melhor performance entre os quatro modelos. A precisão de dados positivo foi de 99%, a sensibilidade de 95% e a acurácia de 98%. O modelo conseguiu prever bem tantos os dados com potencial e sem potencial, o número de falsos foi mínimo frente aos valores verdadeiros (Tabela 7).

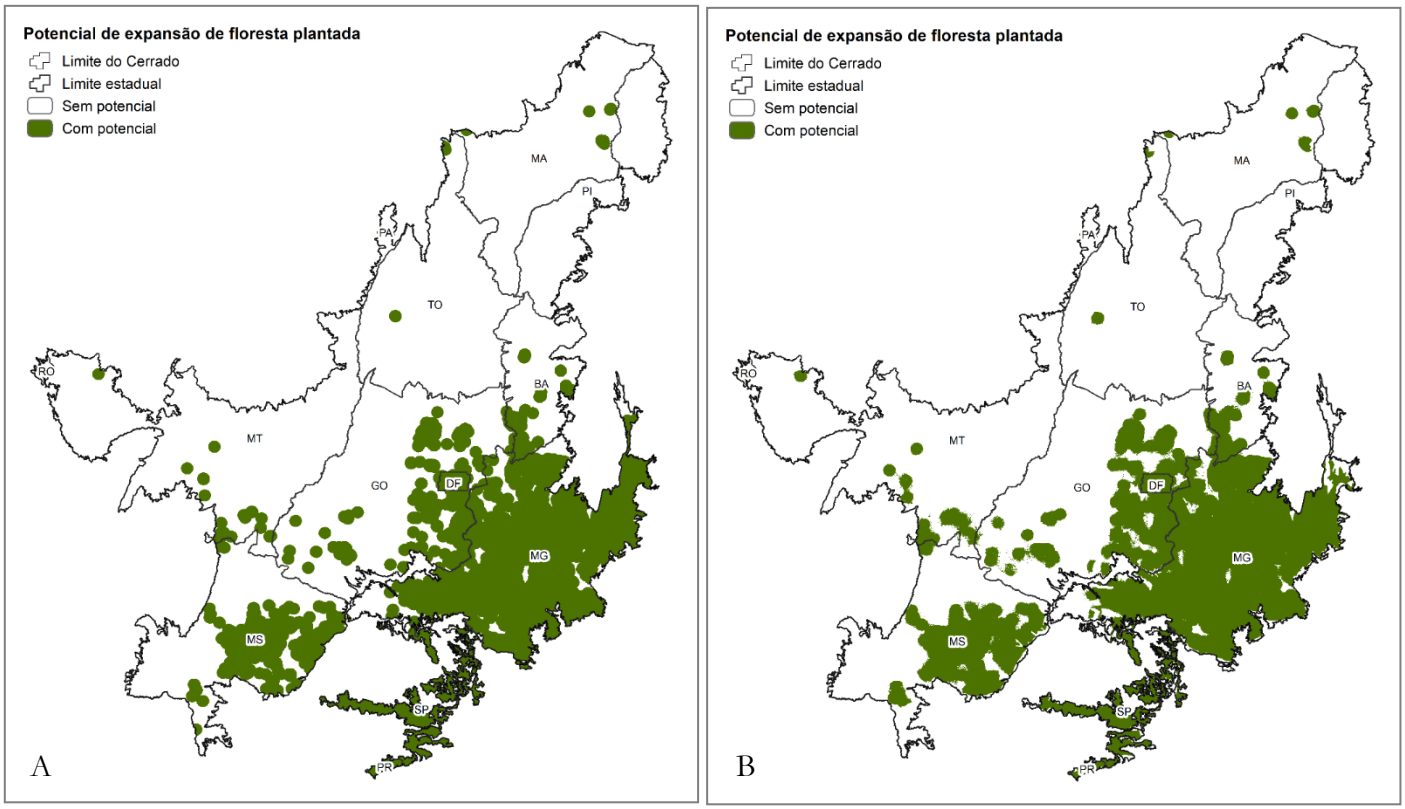
Tabela 7. Matriz de confusão do modelo de expansão da floresta plantada

		Valor real		
		Com potencial	**Sem potencial**	
Valor predito	**Com potencial**	Verdadeiro Positivo	Falso Positivo	Precisão positivo
		130.427	677	99%
	Sem potencial	Falso Negativo	Verdadeiro Negativo	Precisão negativo
		6.257	307.810	98%
		Sensibilidade	Especificidade	Acurácia
		95%	100%	98%

Fonte: Resultados do estudo. Elaboração: Autora.

A boa performance decorre do fato das áreas potenciais reais serem mais contínuas, comparando com os modelos de potencial para expansão da soja e intensificação da pecuária. Isso facilita o reconhecimento de um padrão pelo algoritmo. Na Figura 13 é possível observar essas áreas contínuas e, quando comparado dados reais e preditos, eles são bem semelhantes.

Figura 13. Potencial de expansão da floresta plantada (A) real e (B) predito



Fonte: Resultado do estudo. Elaboração: Autora.

A área com potencial de expansão da floresta plantada predita pelo modelo difere apenas 1,3% (0,7 milhões de hectares) da área real, demonstrando a boa performance do modelo (Tabela 8). As maiores diferenças ocorreram nos estados de Goiás, Minas Gerais e São Paulo.

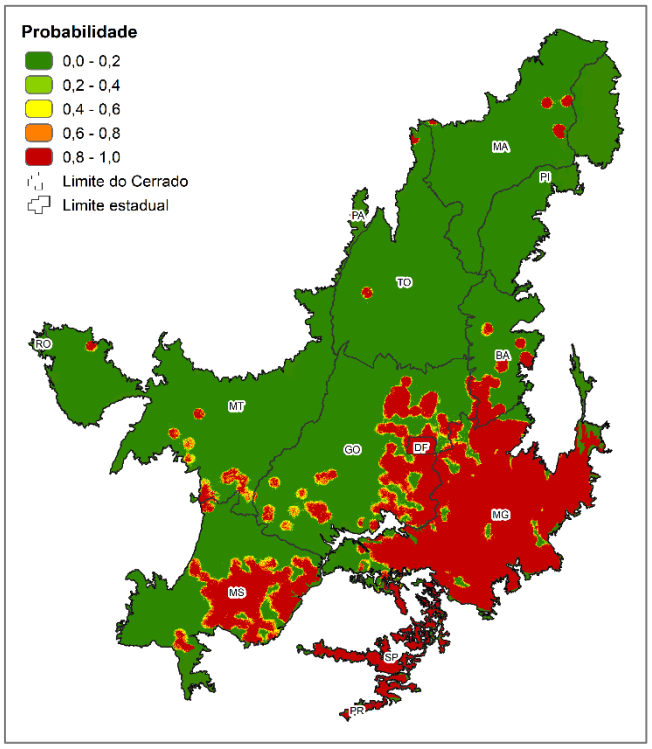
Tabela 8. Área (milhões de hectares) sem e com potencial para expansão da floresta plantada - dados reais e preditos

Estado	Sem potencial		Com potencial	
	Real	Predito	Real	Predito
BA	8,1	8,2	2,2	2,1
DF	0,0	0,0	0,6	0,6
GO	24,4	23,8	9,0	9,7
MA	21,2	21,2	0,5	0,5
MG	3,2	4,1	28,5	27,6
MS	13,7	13,4	8,6	8,8
MT	32,1	31,9	1,6	1,8
PA	0,9	0,9	0,0	0,0
PI	13,3	13,3	0,0	0,0
PR	0,0	0,1	0,3	0,3
RO	0,3	0,1	0,0	0,0
SP	0,2	0,8	4,4	3,7
TO	25,0	25,1	0,3	0,2
TOTAL	**142,5**	**142,7**	**56,0**	**55,3**

Fonte: Resultados do estudo. Elaboração: Autora.

Na Figura 14 é apresentado o mapa de probabilidade do potencial de expansão da floresta plantada. Cerca de 51,6 milhões de hectares tem mais de 80% de probabilidade. Neste caso, o ponto de corte de 50% resultou numa boa performance do modelo.

Figura 14. Mapa de probabilidade do potencial de expansão da floresta plantada



Fonte: Resultado do estudo. Elaboração: Autora.

5.3.4 Modelo final

Ao juntar o resultado dos três modelos acima, ocorreu uma sobreposição de áreas, o que resultou em ter uma mesma área com potencial para mais de uma cultura (soja, pecuária e/ou floresta plantada). E no intuito de ter algo que seja capaz de fazer essa escolha entre os potenciais e evitar a sobreposição, foi desenvolvido o modelo final.

Na Tabela 9 é apresentada a matriz de confusão do modelo final, sendo a acurácia geral do modelo de 80%. A sensibilidade do modelo para a soja foi de 90%, houve bastante certeza quando a predição da área com potencial para soja, por outro lado, o modelo classificou como soja uma quantidade considerável de dados sem potencial, fazendo com que a precisão fosse de 69%.

Para a predição da pecuária, o modelo teve uma boa sensibilidade, de 82%, e uma precisão mediana, de 59%. Houve uma confusão considerável do modelo entre as classes de pecuária e sem potencial, contudo as estatísticas são bem parecidas ao modelo de potencial de intensificação da pecuária.

Como era esperado, a predição da floresta plantada teve um ótimo desempenho, alcançando 99% de sensibilidade e 97% de precisão.

Tabela 9. Matriz de confusão do modelo final

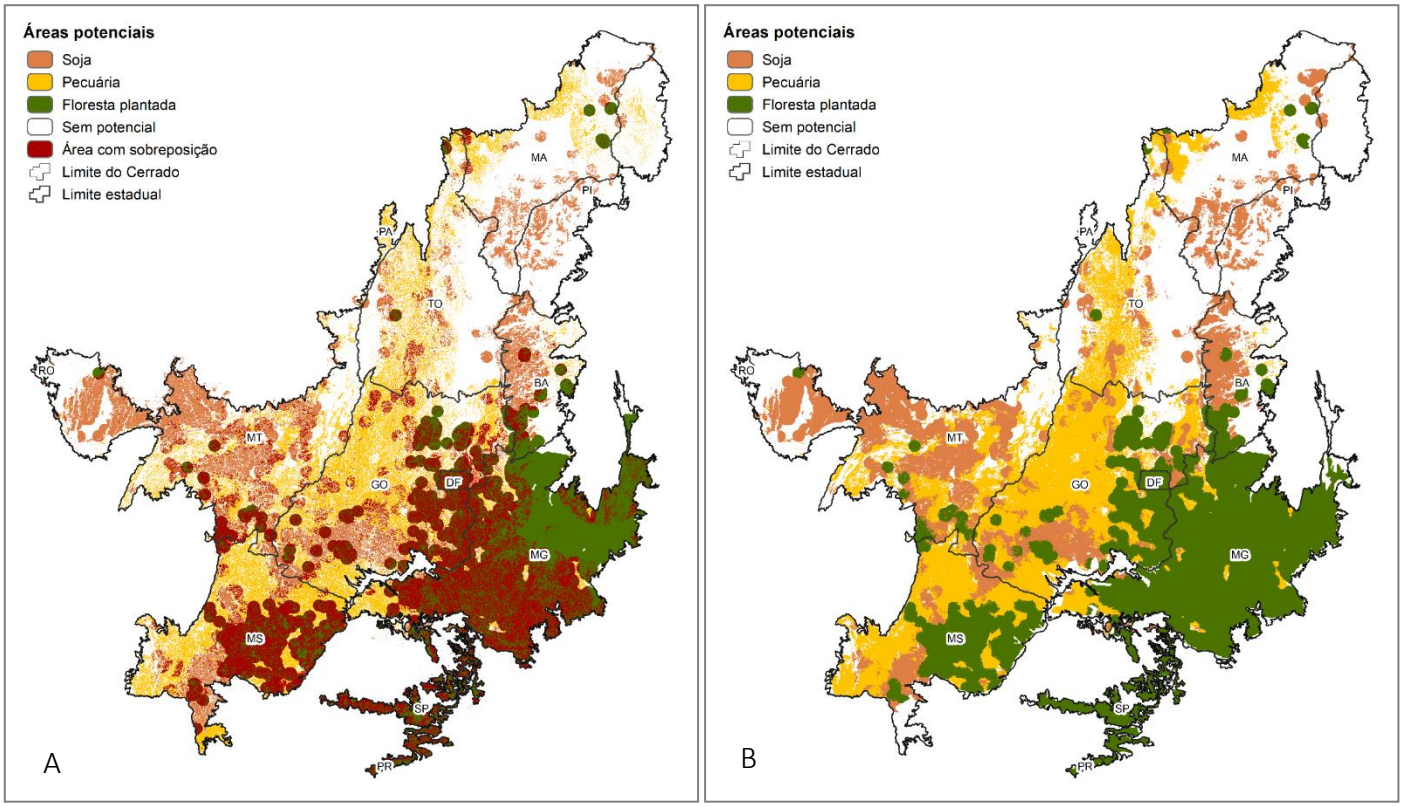
		Valor real				Precisão
		Soja	**Pecuária**	**Floresta**	**Sem potencial**	
Valor predito	**Soja**	50.160	4.422	235	17.629	69%
	Pecuária	3.358	55.362	347	35.522	59%
	Floresta	195	455	63.502	1.128	97%
	Sem potencial	2.239	7.186	56	120.089	93%
		Sensibilidade			Especificidade	Acurácia
		90%	82%	99%	69%	80%

Fonte: Resultado do estudo. Elaboração: Autora.

A Figura 15 apresenta os dados reais e preditos espacializados. No mapa com os dados reais há 5 classes: soja, pecuária, floresta plantada, sem potencial e com sobreposição, já o mapa com os dados preditos apresenta as mesmas classes menos a de sobreposição. A classificação das áreas com potencial para expansão da soja e da intensificação da pecuária compreende grandes áreas, sem muitos detalhes como o dado real, como observado nos modelos anteriores.

Em torno de 35 milhões de hectares apresentou sobreposição de diferentes potenciais, quando analisado os dados reais. Desse total, 26,6 milhões de hectares (75,4%) foram preditos com potencial para expansão da floresta plantada, 5 milhões de hectares (14%) para expansão da soja, 2,5 milhões de hectares (7,2%) para intensificação da pecuária e 1,2 milhão de hectares (3,4%) sem potencial.

Figura 15. Potencial final (A) real e (B) predito



Fonte: Resultados do estudo. Elaboração: Autora.

Usando o modelo final para classificar as áreas de acordo com o potencial, foi obtido como resultado 34,9 milhões de hectares com potencial para expansão da soja, 41,5 milhões com potencial para intensificação da pecuária, 52,4 milhões com potencial para expansão da floresta plantada e 69,3 milhões sem potencial (Tabela 10).

Tabela 10. Área (milhões de hectares) das classes do modelo final

Estado	Soja	Pecuária	Floresta	Sem potencial
BA	3,4	0,2	2,0	4,8
DF	0,0	0,0	0,6	0,0
GO	6,1	15,5	8,7	3,3
MA	3,0	2,6	0,5	15,5
MG	0,4	1,8	27,2	2,4
MS	2,9	8,9	7,9	2,5
MT	14,0	6,3	1,4	12,1
PA	0,0	0,0	0,0	0,8
PI	1,9	0,0	0,0	11,4
PR	0,0	0,0	0,3	0,0
RO	0,0	0,0	0,0	0,0
SP	0,1	0,0	3,7	0,8
TO	3,1	6,3	0,2	15,7
TOTAL	**34,9**	**41,5**	**52,4**	**69,3**

Fonte: Resultado do estudo. Elaboração: Autora.

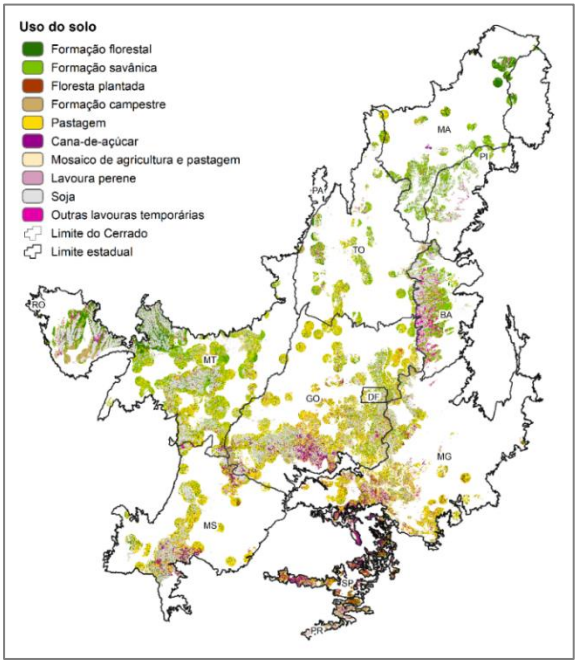
5.4 Aplicações práticas

Ao combinar os dados de uso e cobertura do solo (Mapbiomas – Coleção 5, 2020) com os de potenciais de expansão e intensificação, gerados pelos modelos, foi possível identificar sobre qual uso e ocupação do solo há o potencial para determinada cadeia agropecuária. O potencial em áreas de vegetação pode ser um sinal de alerta, pois a mesma pode ser aberta para produção agrícola, resultando em novos desmatamentos. Já quando o potencial está em áreas abertas, essa expansão/intensificação pode ser incentivada por políticas públicas, diminuindo a pressão de desmatamento das áreas vegetadas.

5.4.1 Soja

A Figura 16 apresenta um mapa que combina as áreas potenciais para expansão da soja e uso do solo. Há potencial tanto em áreas vegetadas, quanto em áreas antrópicas, sendo grande parte em áreas agrícolas como pastagem, lavouras temporárias e áreas de soja.

Figura 16. Uso do solo das áreas com potencial para expansão da soja



Fonte: Resultado do estudo. Elaboração: Autora.

O potencial sobre áreas antrópicas é predominante na região Centro-Sul do Cerrado, que engloba os estados de Minas Gerais, Goiás, Mato Grosso e Mato Grosso do Sul. A região soma 14,7 milhões de hectares de pastagem e 1,7 milhões de lavouras temporárias (exceto soja) com potencial de expansão da soja. Já as áreas com vegetação nativa, estão localizadas predominantemente na região do Matopiba (Maranhão, Tocantins, Piauí e Bahia), somando uma área de 7,4 milhões de hectares com potencial. Os valores mais detalhados são apresentados na Tabela 11.

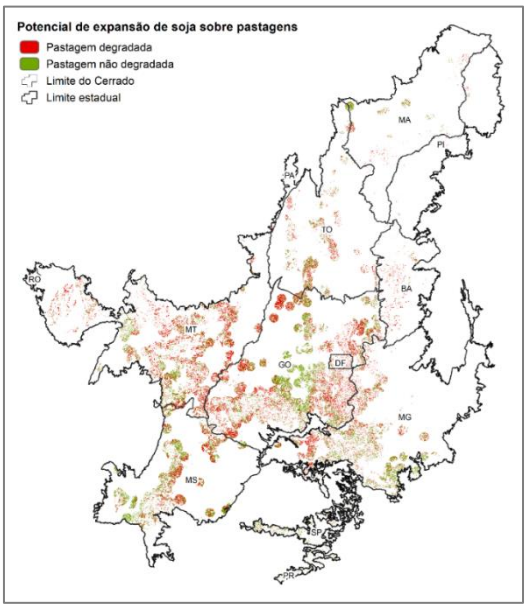
Tabela 11. Área (milhões de hectares) por uso do solo das áreas com potencial para expansão da soja

Estado	Formação florestal	Formação savânica	Floresta plantada	Formação campestre	Pastagem	Soja	Outras lavouras temporárias
BA	0,1	1,5	0,0	0,5	0,2	1,7	0,6
DF	0,0	0,1	0,0	0,1	0,1	0,1	0,0
GO	1,4	2,0	0,1	0,7	5,8	3,6	0,5
MA	0,7	1,3	0,0	0,2	0,3	0,7	0,1
MG	0,6	0,8	0,2	0,3	2,9	1,3	0,4
MS	0,7	0,4	0,1	0,1	2,3	1,6	0,2
MT	2,1	4,2	0,0	0,7	3,3	5,3	0,5
PA	0,0	0,0	0,0	0,0	0,0	0,0	0,0
PI	0,1	0,9	0,0	0,1	0,0	0,8	0,1
PR	0,0	0,0	0,0	0,0	0,0	0,1	0,0
RO	0,0	0,0	0,0	0,0	0,0	0,0	0,0
SP	0,3	0,1	0,3	0,1	0,6	0,3	0,1
TO	0,5	1,3	0,0	0,4	0,9	0,6	0,1
Total	**6,5**	**12,4**	**0,7**	**3,1**	**16,5**	**16,1**	**2,6**

Fonte: Resultado do estudo. Elaboração: Autora.

A área de pastagem com potencial para expansão da soja soma um valor bastante significativo, por esse motivo, foi realizada uma segunda combinação de dados, qualidade de pastagem do LAPIG (2020) e potencial de expansão da soja (Figura 17). Segundo LAPIG (2020), o Cerrado possui uma área de 62,8 milhões de hectares com pastagens, sendo que 23,7 milhões estão degradadas, isso equivale a 38% do total.

Figura 17. Potencial de expansão da soja sobre pastagens degradadas e não degradadas



Fonte: Resultado do estudo. Elaboração: Autora.

No total, há 7,7 milhões de hectares de pastagem degradada e 7,5 milhões de hectares de pastagem não degradada com potencial para expansão da soja. Estas áreas estão concentradas em grande parte nos estados de Goiás, Mato Grosso, Mato Grosso do Sul e Minas Gerais (Tabela 12).

Tabela 12. Área (milhões de hectares) de pastagem com potencial para expansão da soja

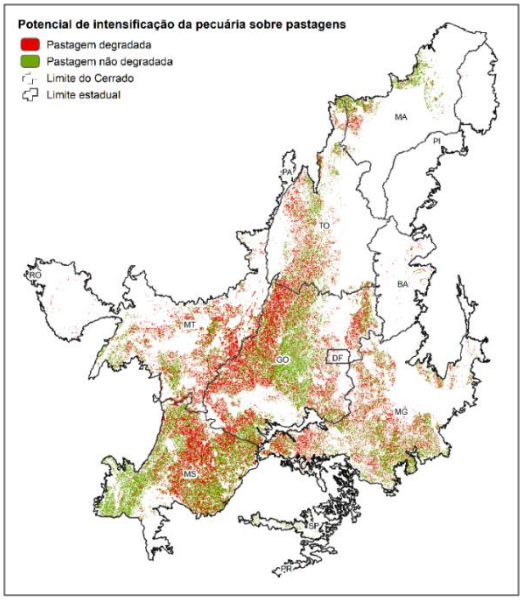
Estado	Pastagem degradada	Pastagem não degradada
BA	0,13	0,05
DF	0,05	0,02
GO	2,79	2,65
MA	0,10	0,13
MG	1,03	1,16
MS	0,88	1,33
MT	2,09	1,48
PA	0,00	0,00
PI	0,02	0,01
PR	0,00	0,01
RO	0,00	0,00
SP	0,06	0,19
TO	0,51	0,47
TOTAL	**7,66**	**7,51**

Fonte: Resultado do estudo. Elaboração: Autora.

5.4.2 Pecuária

Para o modelo de intensificação da pecuária foi analisado o potencial em áreas de pastagem degradada e não degradada (Figura 18).

Figura 18. Potencial de intensificação da pecuária sobre pastagens degradadas e não degradadas



Fonte: Resultado do estudo. Elaboração: Autora.

Dos 62,8 milhões de hectares de pastagem do Cerrado, cerca de 41 milhões tem potencial de intensificação para pecuária, sendo 18 milhões de pastagem degradada e 23 milhões de pastagem não degradada (Tabela 13). Pode ser de interesse a recuperação das pastagens degradadas para intensificação de pecuária, a fim de ter uma produção mais sustentável e liberar áreas para produção de outras culturas, que por sua vez diminui a pressão por novos desmatamentos.

Tabela 13. Área (milhões de hectares) de pastagem com potencial para intensificação da pecuária

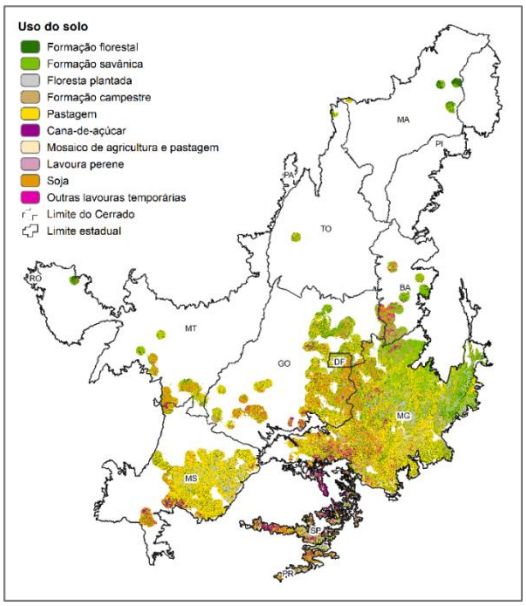
Estado	Pastagem degradada	Pastagem não degradada
BA	2,83	2,35
DF	0,06	0,04
GO	0,03	0,02
MA	5,95	6,83
MG	0,33	0,95
MS	3,83	6,58
MT	3,22	3,91
PA	0,00	0,00
PI	0,00	0,00
PR	0,00	0,00
RO	0,00	0,00
SP	0,04	0,18
TO	1,72	2,05
Total	**18,03**	**22,91**

Fonte: Resultado do estudo. Elaboração: Autora.

5.4.3 Floresta Plantada

A Figura 19 apresenta um mapa que combina as áreas potenciais para expansão da floresta plantada e uso do solo ("Mapbiomas - Coleção 6", 2021). Há potencial tanto em áreas vegetadas, quanto em áreas antrópicas, sendo grande parte em áreas agrícolas como pastagem.

Figura 19. Uso do solo das áreas com potencial para expansão da floresta plantada



Fonte: Resultado do estudo. Elaboração: Autora.

As áreas potenciais estão concentradas nos estados de Minas Gerais, Goiás e Mato Grosso do Sul (Tabela 14). O maior potencial de expansão da floresta plantada está sobre áreas de pastagem (21,5 milhões de hectares), seguido por áreas de formação savânica (11,6 milhões de hectares).

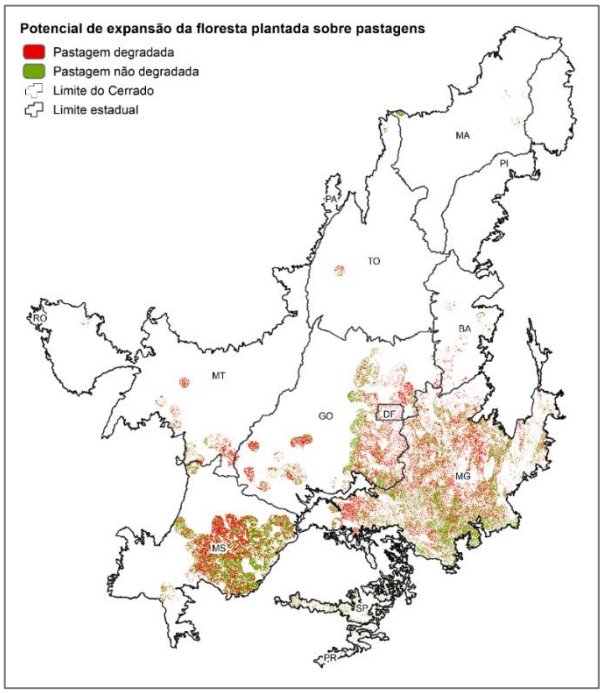
Tabela 14. Área (milhões de hectares) por uso do solo das áreas com potencial para expansão da floresta plantada

Estado	Formação florestal	Formação savânica	Floresta plantada	Formação campestre	Pastagem	Cana-de-açúcar	Soja	Outras lavouras temporárias
BA	0,1	0,9	0,0	0,5	0,1	0,0	0,3	0,2
DF	0,0	0,1	0,0	0,1	0,1	0,0	0,1	0,0
GO	1,2	1,9	0,1	0,7	3,4	0,2	1,8	0,3
MA	0,2	0,2	0,0	0,0	0,1	0,0	0,0	0,0
MG	2,8	7,2	1,6	2,0	10,9	0,5	1,4	0,7
MS	0,9	0,6	0,7	0,2	5,4	0,2	0,5	0,1
MT	0,2	0,5	0,0	0,0	0,5	0,0	0,5	0,1
PA	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0
PI	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0
PR	0,0	0,0	0,1	0,0	0,0	0,0	0,1	0,0
RO	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0
SP	0,4	0,1	0,4	0,1	0,9	1,1	0,3	0,1
TO	0,0	0,1	0,0	0,0	0,1	0,0	0,0	0,0
Total	**6,0**	**11,6**	**2,9**	**3,6**	**21,5**	**1,9**	**5,0**	**1,5**

Fonte: Resultado do estudo. Elaboração: Autora.

Devido ao grande potencial de expansão da floresta plantada sobre pastagem, foi avaliada a qualidade da pastagem nessas áreas, como apresentado na Figura 20.

Figura 20. Potencial de expansão de floresta plantada sobre pastagens degradadas e não degradadas



Fonte: Resultado do estudo. Elaboração: Autora.

No total, há 8,1 milhões de hectares de pastagem degradada e 9,3 milhões de hectares de pastagem não degradada com potencial para expansão da soja. Estas áreas estão concentradas em grande parte nos estados de Goiás, Mato Grosso, Mato Grosso do Sul e Minas Gerais (Tabela 15).

Tabela 15. Área (milhões de hectares) de pastagem com potencial para expansão da floresta plantada

Estado	Pastagem degradada	Pastagem não degradada
BA	0,09	0,04
DF	0,07	0,03
GO	1,63	1,42
MA	0,01	0,05
MG	3,77	4,31
MS	2,12	2,91
MT	0,28	0,18
PA	0,00	0,00
PI	0,00	0,00
PR	0,00	0,01
RO	0,00	0,00
SP	0,08	0,29
TO	0,03	0,02
Total	**8,09**	**9,26**

Fonte: Resultado do estudo. Elaboração: Autora.

6. CONCLUSÕES

Este estudo permitiu concluir que é possível desenvolver modelos preditivos de classificação de potencial do uso do solo para as cadeiras da soja, pecuária e floresta plantada, utilizando ferramentas de aprendizado de máquina.

Foram desenvolvidos quatro modelos utilizando o algoritmo *Random Forest*, entre eles, potencial de expansão da soja, potencial de intensificação da pecuária, potencial de expansão da floresta plantada e, um último modelo, que engloba as três cadeias agropecuárias. Como entradas dos modelos foram utilizados dado climáticos, edáficos, de infraestrutura e socioeconômicos.

O modelo com melhor performance foi o de potencial de expansão da floresta plantada, com 98% de acurácia. Isso se deu pelo fato de as áreas usadas como dado de treino serem bem concentradas e contínuas, o que facilitou o algoritmo desenvolver a lógica entre as variáveis e as classes.

O segundo melhor modelo foi o de potencial de expansão da soja, com uma acurácia de 86%, seguido pelo modelo de potencial de intensificação da pecuária com 79%. O fato das áreas de soja e pastagem, usadas para treinar o algoritmo, serem mais esparsas, principalmente as de pastagens, fizeram com que os modelos classificassem mais áreas com potencial para expansão da soja e intensificação da pastagem do que realmente existe, reduzindo a acurácia de ambos.

Já o modelo final, que reúne as três cadeias agropecuária, teve uma acurácia geral de 80%, que se deve ao fato da boa performance na previsão das classes de floresta planta e sem potencial. Por outro lado, o desempenho para a previsão da classe de soja e pecuária não foram tão boas, chegando a uma acurácia de 69% e 59%, respectivamente.

A aplicação desses modelos identificou um grande potencial de expansão das cadeias agropecuárias sobre áreas antrópicas, principalmente, pastagens degradadas e não degradadas. Essa informação pode ser interessante para guiar possíveis ações de recuperação de áreas degradadas.

De maneira geral, este estudo traz contribuições ao apresentar alternativas para a identificação do potencial de uso do solo, utilizando ferramentas que permite a manipulação de grande quantidade de informação. Os modelos preditivos podem se tornar boas ferramentas para o planejamento estratégico do território. Um destaque deste estudo é que a classificação do potencial foi realizada por pixel com resolução espacial de 1.000 metros, o que permitiu analisar o dado espacializado.

Em trabalhos futuros, é possível desenvolver modelos para outras culturas agrícolas, além de ser expandir para todo o território brasileiro. Também é possível testar diferentes algoritmos de aprendizado de máquina e parâmetros, a fim de ter um modelo com melhor performance.

É de grande valia que, futuramente, tenha um trabalho de longo prazo para coletar dados de campo com informações de áreas que realmente possuem o potencial para cada cadeia agropecuária, para que serviam como dados de treinamento para os modelos. E, dessa forma, gerar um resultado mais próximo com a realidade.

REFERÊNCIAS

ALKIMIM, Akenya Freire De. *University of São Paulo " Luiz de Queiroz " College of Agriculture Multicriteria decision analysis applied to the spatial allocation of crops as a planning support system for agricultural expansion in Brazil*. 2014. 245 f. Universidade de São Paulo, 2014.

BAK, Thomas H.; LENNOX, Graham G. Cross-Validation. *Dementia with Lewy Bodies: and Parkinson's Disease Dementia*, p. 1–8, 2005.

BOLFE, Edson Luis; SANO, Edson Eyji; CAMPOS, Silvia Kanadani. *Dinâmica agrícola no cerrado: análises e projeções*. Brasília: [s.n.], 2020.

CONAB. *Companhia Nacional de Abastecimento*. Disponível em: <https://www.conab.gov.br/armazenagem/sistema-de-cadastro-nacional-de-unidades-armazenadoras-sicarm-1>.

COSTA, Karine *et al. Potencial regional de expansão da soja no Matopiba*. *Paper Knowledge . Toward a Media History of Documents*. São Paulo: [s.n.], 2021.

DAVIS, Jesse; GOADRICH, Mark. The Relationship Between Precision-Recall and ROC Curves Jesse. p. 233–240, 2016. Disponível em: <http://arxiv.org/abs/1609.07195>.

FICK, Stephen E.; HIJMANS, Robert J. WorldClim 2: new 1-km spatial resolution climate surfaces for global land areas. *International Journal of Climatology*, v. 37, n. 12, p. 4302–4315, 2017.

CARNEIRO FILHO, Arnaldo *et al.* Cerrado. Caminhos Para a Ocupação Territorial, Uso Do Solo E Produção Sustentável: Estratégias De Conservação Em Áreas Privadas. 2018.

FILHO, Braz Calderano *et al.* Avaliação da aptidão agrícola das terras da microbacia do córrego da Tábua, no município de Fidélis, RJ. *Boletim de Pesquisa e Desenvolvimento Embrapa*, 2004.

GUIMARÃES, Edson da Silva. *Aprendizado de Máquina aplicado à predição da produtividade da cultura da soja utilizando dados de clima e solo*. 2019. 2019. Disponível em: <https://www.teses.usp.br/teses/disponiveis/55/55137/tde-09062020-123106/>.

HAMADA, Emília; ASSAD, Maria Leonor Lopes; PEREIRA, Danilla Alves. APTIDÃO AGRÍCOLA NA ÁREA DE RECARGA DO AQÜÍFERO GUARANI : CASO DA MICROBACIA HIDROGRÁFICA DO CÓRREGO DO ESPRAIADO. *Engenharia Ambiental*, v. 3, p. 62–71, 2006.

HENGL, Tomislav *et al. SoilGrids250m: Global gridded soil information based on machine learning*. [S.l: s.n.], 2017. v. 12.

HARFUCH, Leila; ROMEIRO, Mariane; PALAURO, Gustavo. *Recuperação de áreas degradadas e reabilitação do solo no Cerrado brasileiro*. . São Paulo: [s.n.]. Disponível em: <https://wwfbr.awsassets.panda.org/downloads/recuperacao_de_areas_degradadas_e_reabilitacao_do_solo_no_cerrado_brasileiro.pdf>. , 2021

HURWITZ, Judith; KIRSCH, Daniel. *Machine Learning for dummies - IBM*. Haboken: John Wiley & Sons, Inc., 2018. v. 35.

IBGE. SIDRA: SISTEMA IBGE DE RECUPERAÇÃO AUTOMÁTICA. *Censo Demográfico 2010: características da população e dos domicílios: resultados do universo*. . Rio de Janeiro: [s.n.], 2011. Disponível em: <http://www.sidra.ibge.gov.br>. Acesso em: 15 jun. 2021.

IBGE. SIDRA: SISTEMA IBGE DE RECUPERAÇÃO AUTOMÁTICA. *Produção da Extração Vegetal e da Silvicultura*. Disponível em: <http://www.sidra.ibge.gov.br>.

IBGE. *Biomas e Sistema Costeiro-Marinho do Brasil: compatível com a escala 1:250.000 [livro eletrônico]*. Rio de Janeiro: [s.n.], 2019. v. 45.

INFRAESTRUTURA, Ministério Da. *Mapas e Bases dos Modos de Transportes*. Disponível em: <http://www.infraestrutura.gov.br/component/content/article/63-bit/5124-bitpublic.html#maprodo>. Acesso em: 17 maio 2020.

LAPIG. *Digital Atlas of Brazilian Pastureland*. Disponível em: <https://pastagem.org/index.php/pt-br/tools/atlas-digital-das-pastagens-brasileiras>. Acesso em: 10 abr. 2020.

LIAKOS, Konstantinos G. *et al.* Machine learning in agriculture: A review. *Sensors (Switzerland)*, v. 18, n. 8, p. 1–29, 2018.

LORENSINI, Carolina Lobello. *Metodologia para classificação da aptidão agrícola de municípios*. 2019. 2019.

LUMBRERAS, José Francisco *et al.* Aptidão Agrícola das Terras do Matopiba. *Documentos*, v. 179, p. 49, 2015. Disponível em: <https://www.embrapa.br/busca-de-publicacoes/-/publicacao/1025303/aptidao-agricola-das-terras-do-matopiba>.

MapBiomas - Coleção 5. Disponível em: <20/09/2020>.

Mapbiomas - Coleção 6. Disponível em: <https://mapbiomas.org/>.

MENDAS, Abdelkader; DELALI, Amina. Integration of MultiCriteria Decision Analysis in GIS to develop land suitability for agriculture: Application to durum wheat cultivation in the region of Mleta in Algeria. *Computers and Electronics in Agriculture*, v. 83, p. 117–126, 2012. Disponível em: <http://dx.doi.org/10.1016/j.compag.2012.02.003>.

MINSKY, Marvin. *Semantic information processing*. [S.l.]: Cambridge, Mass., MIT Press, 1968.

MIRANDA, E. E. De. *Brasil em Relevo*. . Campinas: Embrapa Monitoramento por Satélite. Disponível em: <http://www.relevobr.cnpm.embrapa.br/>. , 2005

PAM/IBGE. *Pesquisa Agrícola Municipal*. Disponível em: <https://sidra.ibge.gov.br/tabela/1612>. Acesso em: 5 jan. 2020.

PEDREGOSA, Fabian *et al.* Scikit-learn: Machine Learning in Python. *Environmental Health Perspectives*, v. 12, n. 85, p. 2825–2830, 2011.

PPM/IBGE. *Pesquisa Pecuária Municipal*. Disponível em: <https://sidra.ibge.gov.br/pesquisa/ppm/quadros/brasil/2018>. Acesso em: 20 mar. 2020.

RAMALHO-FILHO, Antonio; BEEK, K J. *Sistema de avaliação da aptidão agrícola das terras*. 3. ed. Rio de Janeiro: EMBRAPA-

CNPPS, 1995.

ROMEIRO, Mariane *et al.* Caminhos para a ocupação territorial, uso do solo e produção sustentável. *Iniciativa para o Uso da Terra - INPUT*, 2018. Disponível em: <https://www.inputbrasil.org/wp-content/uploads/2018/06/CERRADO-CAMINHOS-PARA-OCUPACAO-TERRITORIAL-SUSTENTAVEL-EXPANSÃO-DA-SOJA-FINAL.pdf>.

RUDORFF, Bernardo; RISSO, Joel. *Análise Geoespacial da Dinâmica das Culturas Anuais no Bioma Cerrado: 2000 a 2014*. . Florianópolis: [s.n.], 2015.

SAMUEL, Arthur L. Some Studies in Machine Learning Using the Game of Checkers. *IBM Journal*, p. 210–229, 1959.

SANTANA, Rodrigo. *Validação Cruzada: Aprenda de forma simples como usar essa técnica*. Disponível em: <https://minerandodados.com.br/validacao-cruzada-aprenda-de-forma-simples-como-usar-essa-tecnica/>.

SHALEV-SHWARTZ, Shai; BEN-DAVID, Shai. *Understanding machine learning: From theory to algorithms*. [S.l: s.n.], 2013. v. 9781107057.

SHALEV-SHWARTZ, Shai; BEN-DAVID, Shai. *Understanding machine learning: From theory to algorithms*. 1. ed. New York: Cambridge University Press, 2014. v. 9781107057. Disponível em: <cs.huji.ac.il/~shais/UnderstandingMachineLearning/understanding-machine-learning-theory-algorithms.pdf>.

SPAROVEK, Gerd *et al.* A revisão do Código Florestal brasileiro. *Novos Estudos - CEBRAP*, n. 89, p. 111–135, mar. 2011. Disponível em: <http://www.scielo.br/scielo.php?script=sci_arttext&pid=S0101-33002011000100007&lng=pt&tlng=pt>.

VALERIANO, Márcio De Morisson. Topodata : Guia Para Utilização De Dados. *Inpe*, v. 8, p. 73, 2008.

VIEIRA, Fábio D.; OLIVEIRA, Stanley R.De M.; PAIVA, Samuel R. Metodologia baseada em técnicas de mineração de dados para suporte à certificação de raças de ovinos. *Engenharia Agricola*, v. 35, n. 6, p. 1172–1186, 2015.

WATANABE, Hiroki. Cross-Validation. *Contemporary Interventional Ultrasonography in Urology*, p. 1–6, 2009.