

**UNIVERSIDADE FEDERAL DE MINAS GERAIS  
INSTITUTO DE GEOCIÊNCIAS  
PROGRAMA DE PÓS-GRADUAÇÃO EM GEOGRAFIA**

Carlos Eduardo Fernandes de Holanda

**AVALIAÇÃO DO DESEMPENHO ESTATÍSTICO DOS ALGORITMOS DE  
CLASSIFICAÇÃO RANDOM FOREST (RF), DECISION TREE (DT) E SUPPORT  
VECTOR MACHINE (SVM) PARA IMAGENS DE SATÉLITES COM DIFERENTES  
CUBOS DE DADOS: estudo de caso no bioma cerrado.**

Belo Horizonte  
2021

Carlos Eduardo Fernandes de Holanda

**AVALIAÇÃO DO DESEMPENHO ESTATÍSTICO DOS ALGORITMOS DE CLASSIFICAÇÃO RANDOM FOREST (RF), DECISION TREE (DT) E SUPPORT VECTOR MACHINE (SVM) PARA IMAGENS DE SATÉLITES COM DIFERENTES CUBOS DE DADOS: estudo de caso no bioma cerrado.**

Dissertação apresentada ao Programa de Pós-Graduação do Departamento de Geografia do Instituto de Geociências da Universidade Federal de Minas Gerais como requisito parcial para obtenção do título de mestre em Geografia.

Área de concentração: Análise Ambiental

Linha de pesquisa: Geografia aplicada e geotecnologias

Orientador: Dr. Diego Rodrigues Macedo

Coorientador: Dr. Rodrigo Affonso de Albuquerque Nóbrega

Belo Horizonte  
2021

H722a Holanda, Carlos Eduardo Fernandes de.  
2021 Avaliação do desempenho estatístico dos algoritmos de classificação Random Forest (RF), Decision Tree (DT) e Support Vector Machine (SVM) para imagens de satélite com diferentes cubos de dados [manuscrito] : estudo de caso no bioma cerrado / Carlos Eduardo Fernandes de Holanda. – 2021. 123 f., enc.: il. (principalmente color.)

Orientador: Diego Rodrigues Macedo.  
Coorientador: Rodrigo Affonso de Albuquerque Nóbrega.  
Dissertação (mestrado) – Universidade Federal de Minas Gerais, Instituto de Geociências, 2021.

Área de concentração: Análise Ambiental.

Linha de pesquisa: Geografia Aplicada e Geotecnologias.

Bibliografia: f. 117-123.

1. Aprendizado do computador – Teses. 2. Solo – Uso – Teses. 3. Mapeamento da cobertura do solo – Teses. 4. Sensoriamento remoto – Teses. I. Macedo, Diego Rodrigues. II. Nóbrega, Rodrigo Affonso de Albuquerque. III. Universidade Federal de Minas Gerais. Instituto de Geociências. IV. Título.

CDU: 39(81)(=98):801.563.1



UNIVERSIDADE FEDERAL DE MINAS GERAIS  
INSTITUTO DE GEOCIÊNCIAS  
COLEGIADO DO CURSO DE PÓS-GRADUAÇÃO EM GEOGRAFIA

### FOLHA DE APROVAÇÃO

**Avaliação do desempenho estatístico dos algoritmos de  
classificação Random Forest (RF), Decision Tree (DT) e Support  
Vector Machine (SVM) para imagens de satélite com diferentes  
cubos de dados: Estudo de caso no bioma cerrado**

**CARLOS EDUARDO FERNANDES DE HOLANDA**

Dissertação submetida à Banca Examinadora designada pelo Colegiado do Programa de Pós-graduação em GEOGRAFIA, como requisito para obtenção do grau de Mestre em GEOGRAFIA, área de concentração ANÁLISE AMBIENTAL.

Aprovada em 30 de novembro de 2021, pela banca constituída pelos membros:

Prof. Glauco José de Matos Umbelino

UFVJM

Profa. Dra. Úrsula Ruchkys de Azevedo

UFMG

Prof. Dr. Rodrigo Affonso de Albuquerque Nóbrega

UFMG

Prof. Diego Rodrigues Macedo - Orientador

UFMG

Belo Horizonte, 30 de novembro de 2021.



Documento assinado eletronicamente por **Rodrigo Affonso de Albuquerque Nobrega, Professor do Magistério Superior**, em 30/11/2021, às 11:05, conforme horário oficial de Brasília, com fundamento no art. 5º do [Decreto nº 10.543, de 13 de novembro de 2020](#).



Documento assinado eletronicamente por **Ursula Ruchkys de Azevedo, Chefe de departamento**, em 30/11/2021, às 11:05, conforme horário oficial de Brasília, com fundamento no art. 5º do [Decreto nº 10.543, de 13 de novembro de 2020](#).

---



Documento assinado eletronicamente por **Diego Rodrigues Macedo, Professor do Magistério Superior**, em 30/11/2021, às 11:06, conforme horário oficial de Brasília, com fundamento no art. 5º do [Decreto nº 10.543, de 13 de novembro de 2020](#).

---



Documento assinado eletronicamente por **Glauco José de Matos Umbelino, Usuário Externo**, em 30/11/2021, às 14:10, conforme horário oficial de Brasília, com fundamento no art. 5º do [Decreto nº 10.543, de 13 de novembro de 2020](#).

---



A autenticidade deste documento pode ser conferida no site [https://sei.ufmg.br/sei/controlador\\_externo.php?acao=documento\\_conferir&id\\_orgao\\_acesso\\_externo=0](https://sei.ufmg.br/sei/controlador_externo.php?acao=documento_conferir&id_orgao_acesso_externo=0), informando o código verificador **1112625** e o código CRC **363294EC**.

---

A todos que de certa forma acreditaram em meu potencial. Em especial aos meus pais, Darci e Angela, à minha esposa, Stephanie e principalmente a Deus por permitir mais este período na vida.

**Dedico.**

## **AGRADECIMENTOS**

Agradeço aos meus pais e irmãos pelo apoio e incentivo em todas as etapas da minha educação, que além de me oferecerem a oportunidade de estudar sempre estiveram presentes participando de todas as minhas decisões.

À UFMG e principalmente ao IGC, seus professores e funcionários pelo acolhimento e oportunidade de continuar aprendendo.

Agradeço ao meu orientador prof. Dr. Diego Rodrigues Macedo por aceitar me orientar num momento decisivo do mestrado, pela paciência e disponibilidade que foram primordiais na elaboração deste trabalho, pelos grandes ensinamentos, incentivos, liberdade e confiança nesta pesquisa.

Agradeço também ao meu coorientador prof. Dr. Rodrigo Affonso de Albuquerque Nóbrega pela disponibilidade e auxílio nas discussões conceituais e práticas, além da compreensão referente ao tempo para a elaboração do trabalho.

Agradeço à Prefeitura Municipal de Santa Luzia/MG, em especial a Secretaria de Desenvolvimento Urbano e Habitação, pela disponibilidade e entendimento de todos.

Agradeço à minha esposa Stephanie Rocha Holanda, pelos incentivos, calma e amor nos períodos mais turbulentos.

Agradeço a CAPES pelo auxílio financeiro concedido nos primeiros seis meses do mestrado e ao Projeto PROECOS (P&D Aneel GT599) o qual esta pesquisa se insere.

Agradeço aos professores Dr. Erivaldo Antonio da Silva, Dra. Úrsula Ruchkys de Azevedo e Dr. Glauco José de Matos Umbelino que aceitaram prontamente o convite para participar da banca de seminário e defesa da dissertação e pelos direcionamentos apontados.

## Resumo

O principal objetivo deste trabalho é analisar e comparar o desempenho de três algoritmos baseados em aprendizado de máquina: Random Forest (RF), Decision Tree (DT) e Support Vector Machine (SVM) na classificação de imagens de satélite de uso e cobertura da terra utilizando informações altimétricas em diferentes cubos de dados. Nesse sentido, foram interpretados os resultados encontrados nos índices estatísticos de validação e na concordância entre as classes na matriz de confusão, o que pode auxiliar estudantes e pesquisadores da área de sensoriamento remoto a escolherem o algoritmo de classificação que melhor se adequa às suas pesquisas. Para este estudo foram utilizadas imagens do satélite Sentinel 2 e outra imagem que empilha as do Sentinel 2 com imagens derivadas da missão Shuttle Radar Topography Mission (SRTM) no intuito de identificar e interpretar as diferenças nas classificações dessas imagens. O classificador que obteve os melhores desempenhos foi o SVM, tanto para as imagens com os dados altimétricos quanto para as imagens sem estas informações complementares. Apesar do DT ter apresentado resultados satisfatórios, mas inferiores ao SVM, o classificador de DT obteve tempo de processamento para a classificação de imagens muito menor que o do SVM. Diante desse dado, observa-se que, caso a área de pesquisa seja muito extensa ou tenha várias pequenas áreas aliadas a um computador com processador não adequado, o classificador SVM não seria a melhor opção. O algoritmo RF obteve praticamente o mesmo tempo de processamento que o DT, mas atingiu os menores índices estatísticos entre os três classificadores. Ainda, verificou-se que a utilização do cubo de dados com a imagem do SRTM e suas derivações como declividade e rugosidade na classificação do uso e cobertura da terra para os três algoritmos mostraram resultados superiores em comparação com as imagens sem estes dados altimétricos.

**Palavras-chave:** Aprendizado de Máquina (ML). Uso e cobertura da terra. Dados altimétricos. Sensoriamento Remoto.



## **Abstract**

*The main objective of this work is to analyze and compare the performance of three algorithms based on machine learning: Random Forest (RF), Decision Tree (DT) and Support Vector Machine (SVM) in the classification of land use and land cover satellite images using altimetric information in different data cubes. In this sense, the results found in the statistical indices of validation and agreement between classes in the confusion matrix were interpreted, which can help students and researchers in the field of remote sensing to choose the classification algorithm that best suits their research. For this study, images from the Sentinel 2 satellite and another image that stacks those from Sentinel 2 with images derived from the Shuttle Radar Topography Mission (SRTM) were used in order to identify and interpret the differences in the classifications of these images. The classifier that obtained the best performance was the SVM, both for images with altimetric data and for images without this additional information. Although the DT presented satisfactory results, but inferior to the SVM, the DT classifier had much lower processing time for the classification of images than the SVM. Given this data, it is observed that, if the research area is very extensive or has several small areas combined with a computer with an unsuitable processor, the SVM classifier would not be the best option. The RF algorithm obtained practically the same processing time as the DT, but reached the lowest statistical indexes among the three classifiers. Furthermore, it was found that the use of the data cube with the SRTM image and its derivations such as slope and roughness in the classification of land use and land cover for the three algorithms showed superior results compared to the images without these altimetric data.*

**Keywords:** Machine Learning (ML). Land use and land cover. Altimetric data. Remote sensing.

## LISTA DE FIGURAS

<b>Figura 1:</b> Representação do método de Árvore de Decisão .....	22
<b>Figura 2:</b> Representação do método de Florestas Aleatórias .....	24
<b>Figura 3:</b> Representação do método de Máquina de Vetores de Suporte .....	26
<b>Figura 4:</b> Mapa de localização da área de estudo .....	31
<b>Figura 5:</b> Fluxograma de atividades desenvolvidas no estudo .....	40
<b>Figura 6:</b> Grade de distribuição dos dados do SRTM .....	44
<b>Figura 7:</b> Grade das cenas do Sentinel 2 .....	45
<b>Figura 8:</b> Bandas utilizadas do satélite Sentinel 2 .....	46
<b>Figura 9:</b> Chave de Interpretação das classes da imagem Sentinel 2 e fotos do Google .....	48
<b>Figura 10:</b> Mapa das Áreas Urbanas vetorizadas manualmente .....	49
<b>Figura 11:</b> Imagem da área de estudo para a altimetria .....	50
<b>Figura 12:</b> Imagem da área de estudo para a declividade .....	51
<b>Figura 13:</b> Imagem da área de estudo para a rugosidade .....	52
<b>Figura 14:</b> Imagem mosaicada e empilhada do Sentinel 2, SRTM, declividade e rugosidade .....	53
<b>Figura 15:</b> Fluxograma da classificação e pós-classificação das imagens no <i>Orfeo ToolBox</i> .....	54
<b>Figura 16:</b> Mapa de classificação sem altimetria – Máquina de Vetores de Suporte (SVM) .....	68
<b>Figura 17:</b> Mapa de classificação com altimetria – Máquina de Vetores de Suporte (SVM) .....	74
<b>Figura 18:</b> Mapa de classificação sem altimetria – Árvore de Decisão (DT) .....	80
<b>Figura 19:</b> Mapa de classificação com altimetria – Árvore de Decisão (DT) .....	86
<b>Figura 20:</b> Mapa de classificação sem altimetria – Floresta Aleatória (RF) .....	91
<b>Figura 21:</b> Mapa de classificação com altimetria – Floresta Aleatória (RF) .....	98
<b>Figura 22:</b> Mapa classificação (SVM) com Sentinel-2 composição colorida .....	107
<b>Figura 23:</b> Mapa classificação (SVM) com Sentinel-2 composição falsa cor .....	108
<b>Figura 24:</b> Mapa de classificação da ESRI para o ano de 2020 .....	109
<b>Figura 25:</b> Mapa de classificação (SVM) com as classes da ESRI .....	112

## LISTA DE TABELAS

<b>Tabela 1:</b> Caracterização matemática da matriz de confusão .....	27
<b>Tabela 2:</b> Avaliação do grau de concordância do índice <i>kappa</i> .....	29
<b>Tabela 3:</b> Dados dos municípios que englobam a bacia hidrográfica .....	32
<b>Tabela 4:</b> Dados sobre o satélite Sentinel 2 e o sensor utilizado .....	42
<b>Tabela 5:</b> Parâmetros para o treinamento das imagens .....	55
<b>Tabela 6:</b> Parâmetros para a pós-classificação das imagens .....	57
<b>Tabela 7:</b> Parâmetros para o algoritmo <i>Random Forest</i> (RF) .....	58
<b>Tabela 8:</b> Parâmetros para o algoritmo <i>Decision Tree</i> (DT) .....	59
<b>Tabela 9:</b> Parâmetros para o algoritmo <i>Support Vector Machine</i> (LibSVM) .....	60
<b>Tabela 10:</b> Quadro das amostras utilizadas no treinamento dos classificadores mostrando estatísticas globais .....	62
<b>Tabela 11:</b> Quadro comparativo dos resultados <i>kappa</i> nos classificadores .....	65
<b>Tabela 12:</b> Quadro comparativo do tempo gasto do processamento dos dados de treinamento, classificação e pós-classificação .....	65
<b>Tabela 13:</b> Matriz de confusão (pixels), erro de omissão e comissão, exatidão global e índice <i>kappa</i> para a Máquina de Vetores de Suporte (SVM) sem altimetria .....	70
<b>Tabela 14:</b> Matriz de confusão (pixels), erro de omissão e comissão, exatidão global e índice <i>kappa</i> para a Máquina de Vetores de Suporte (SVM) com altimetria .....	76
<b>Tabela 15:</b> Matriz de confusão (pixels), erro de omissão e comissão, exatidão global e índice <i>kappa</i> para a Árvore de Decisão (DT) sem altimetria .....	82
<b>Tabela 16:</b> Matriz de confusão (pixels), erro de omissão e comissão, exatidão global e índice <i>kappa</i> para a Árvore de Decisão (DT) com altimetria .....	88
<b>Tabela 17:</b> Matriz de confusão (pixels), erro de omissão e comissão, exatidão global e índice <i>kappa</i> para a Floresta Aleatória (RF) sem altimetria .....	94
<b>Tabela 18:</b> Matriz de confusão (pixels), erro de omissão e comissão, exatidão global e índice <i>kappa</i> para a Floresta Aleatória (RF) com altimetria .....	100
<b>Tabela 19:</b> Quadro comparativo dos resultados das matrizes de confusão dos algoritmos SVM, DT e RF com e sem os dados altimétricos .....	104
<b>Tabela 20:</b> Quadro comparativo final dos resultados <i>kappa</i> e tempo de processamento obtidos nos classificadores abordados e no mapa da ESRI .....	106

## LISTA DE ABREVIATURAS

ANEEL - Agência Nacional de Energia Elétrica  
CEMIG - Companhia Energética de Minas Gerais  
DT - Decision Tree  
ESA - European Space Agency  
ESRI - Environmental Systems Research Institute  
IBGE - Instituto Brasileiro de Geografia e Estatística  
IC - Inteligência Computacional  
IEF - Instituto Estadual de Florestas  
MDE - Modelo Digital de Elevação  
ML - Machine Learning  
MMA - Ministério do Meio Ambiente  
MSI - Multispectral Imaging  
NASA - National Aeronautics and Space Administration  
NGA - National Geospatial-Intelligence Agency  
OTB - Orfeo Toolbox  
PDI - Processamento Digital de Imagens  
PROECOS - Processos ecológicos  
RADAR - Radio Detection and Ranging  
RF - Random Forest  
SAR - Synthetic Aperture Radar  
SIG - Sistemas de Informações Geográficas  
SNAP - Sentinel Application Platform  
SR - Sensoriamento Remoto  
SRTM - Shuttle Radar Topography Mission  
SVM - Support Vector Machine  
TRI - Terrain Ruggedness Index  
UC - Unidade de Conservação  
USGS - United States Geological Survey  
WWF - World Wide Fund for Nature

# SUMÁRIO

<b>1 INTRODUÇÃO .....</b>	<b>11</b>
1.1 Objetivos .....	17
1.1.1 Objetivo Geral .....	17
1.1.2 Objetivos específicos .....	17
<b>2 BASES CONCEITUAIS .....</b>	<b>18</b>
2.1 Sensoriamento Remoto e Processamento Digital de Imagens no mapeamento do uso e cobertura da terra .....	18
2.1.2 Cubo de dados .....	19
2.1.3 Classificadores .....	20
2.2 Validação da classificação e coeficientes estatísticos .....	26
<b>3 CARACTERIZAÇÃO DA ÁREA DE ESTUDO .....</b>	<b>30</b>
3.1 Cobertura Vegetal .....	35
3.2 Macrodrenagem da área de estudo .....	37
<b>4 METODOLOGIA .....</b>	<b>37</b>
4.1 Recursos materiais .....	38
4.2 Pré-processamento das imagens de Declividade e Rugosidade .....	41
4.3 Aquisição de imagens .....	42
4.4 Seleção de amostras .....	46
4.5 Processamento das imagens SRTM, Declividade e Rugosidade .....	49
4.5.1 Imagem da área de estudo para o SRTM .....	49
4.5.2 Imagem da área de estudo para a declividade .....	50
4.5.3 Imagem da área de estudo para a rugosidade .....	51
4.5.4 Imagem empilhada para o Sentinel 2 .....	52
4.6 Classificação .....	53
4.7 Classificadores .....	57
4.8 Mapa global de uso e cobertura da terra da ESRI para o ano de 2020 .....	61
<b>5 RESULTADOS E DISCUSSÕES .....</b>	<b>62</b>
5.1 Avaliação dos Classificadores .....	65
5.1.1 Imagem classificada sem altimetria para o algoritmo de Support Vector Machine (SVM) .....	67

5.1.2 Imagem classificada com altimetria para o algoritmo de Support Vector Machine (SVM) .....	73
5.1.3 Imagem classificada sem altimetria para o algoritmo de Árvore de Decisão (DT) .....	79
5.1.4 Imagem classificada com altimetria para o algoritmo de Árvore de Decisão (DT) .....	85
5.1.5 Imagem classificada sem altimetria para o algoritmo de Floresta Aleatória (RF) .....	91
5.1.6 Imagem classificada com altimetria para o algoritmo de Floresta Aleatória (RF) .....	97
5.1.7 Comparação dos resultados das matrizes de confusão dos algoritmos SVM, DT e RF com e sem os dados altimétricos .....	103
5.1.8 Mapa das áreas classificadas (SVM altimetria) com a realidade de campo – Sentinel 2 .....	106
5.2 Mapa de uso e cobertura da terra da ESRI para o ano de 2020 .....	108
<b>6 CONCLUSÃO .....</b>	<b>113</b>
<b>REFERÊNCIAS .....</b>	<b>117</b>

## 1 INTRODUÇÃO

Os efeitos das ações antrópicas sobre o meio ambiente provocam mudanças significativas no equilíbrio e na dinâmica da natureza. O crescimento populacional, evidenciado em sua maior parte nas áreas urbanas, necessita cada vez mais de recursos, como água potável e alimentos. Essa pressão sobre o meio ambiente na procura de espaços geográficos para o cultivo e de melhorias nas estratégias de produção de alimentos vem cobrando dos planejadores que sejam formuladas políticas públicas que garantam não somente sustentabilidade socioeconômica, mas também ambiental (AHLERT, 2007).

O Brasil é composto por seis biomas de características diferentes: Amazônia, Mata Atlântica, Cerrado, Caatinga, Pampa e Pantanal. Com as maiores dimensões em área, destacam-se a Amazônia e o Cerrado, ocupando 49% e 24% de todo o território nacional, respectivamente (IBGE, 2020). Entre as demandas que recaem sobre o bioma Cerrado, duas configuram-se pelas principais atividades econômicas, atualmente apoiadas pelas políticas agrícolas, que estão ameaçando a sua biodiversidade: a monocultura de grãos e a pecuária de baixa tecnologia (IEF, 2020). Além disso, nas últimas décadas vem ocorrendo um rápido processo de urbanização e expansão de atividades agropecuárias, com estimativa atual de desmatamento no Cerrado chegando a três milhões de hectares por ano no país (IEF, 2020). A mineração e o aumento da população também se apresentam como dois outros fatores que se somam às pressões antrópicas sobre o bioma (WWF, 2012).

Nesse âmbito, cumpre lembrar que o bioma Cerrado ocupa uma área de 2.036.448 km<sup>2</sup> (MMA, 2020) e apresenta abundância de biodiversidade de espécies endêmicas, sendo reconhecido como a savana mais rica do mundo do ponto de vista da biodiversidade. Em termos numéricos, ele abriga 11.627 espécies de plantas nativas e mais de 2.550 espécies de animais (MMA, 2020), ultrapassando a marca das 4.800 espécies de plantas e vertebrados endêmicos (STRASSBURG *et al.*, 2017). Ao lado da Mata Atlântica, o Cerrado é considerado como um dos *hotspots* mundiais de biodiversidade, o que significa que esses biomas reúnem as condições de possuir alta biodiversidade e alto grau de ameaça de degradação (MYERS *et al.*, 2000).

Dentro do Estado de Minas Gerais, o Cerrado conta com uma área de 31.119,87 km<sup>2</sup>, o que corresponde a 10,3% da área original de cerca de 300 mil km<sup>2</sup> no Estado (IEF, 1996). Tais áreas estão divididas entre as regiões do Alto e Médio São Francisco, Alto e Médio Jequitinhonha, Campo das Vertentes, Triângulo e Alto Paranaíba e Zona Metalúrgica, abrangendo relevo plano a suavemente ondulado (BRANDÃO, 2000). O Cerrado abrange também três das maiores bacias hidrográficas da América do Sul, contribuindo com 43% das águas superficiais do Brasil fora da região amazônica (STRASSBURG *et al.*, 2017).

A construção de estradas, a urbanização e o desmatamento desenfreado, devido ao avanço da agricultura e da pecuária, ocasionaram a fragmentação dos habitats, tornando-os incapazes de proteger sua biodiversidade – fator este que mais contribui para a degradação do Cerrado (MACHADO *et al.*, 2004; TEIXEIRA *et al.*, 2004). De acordo com o *World Wide Fund For Nature* (2012), entre os anos de 2002 a 2011, a fragmentação de habitats e a destruição representaram em torno de 66% da área total do Cerrado, sendo destes 60% destinados à pecuária e 6% aos grãos, dos quais 1% apenas estava voltado para a produção de soja.

Nesse sentido, a pressão acentuada pela expansão da agricultura e pela proteção limitada conseguem explicar as projeções de que os atuais 31-34% das áreas remanescentes do Cerrado provavelmente sejam dizimados até o ano de 2050 (STRASSBURG *et al.*, 2017). Essa projeção pode ser explicada pela retirada parcial ou total da cobertura vegetal tem gerado reduções do acúmulo da produção de biomassa vegetal, com significativas perdas da cobertura dos solos (COSTA *et al.*, 2009).

Nesse contexto, o monitoramento e as análises da cobertura vegetal por meio de aplicações do sensoriamento remoto (SR) aparecem como ações que se configuram importantes ferramentas para o levantamento das aceleradas mudanças que ocorrem na superfície terrestre (MARCUSSEI *et al.*, 2010). Dentre as várias aplicações em que se insere o SR, estão: a) a estimativa da biomassa e da produtividade da cultura; b) o acompanhamento do estresse hídrico e da robustez das plantas; e c) a verificação do estágio fenológico (BRANDÃO, 2009).

Para compreender melhor o SR, o seu desenvolvimento histórico pode ser dividido em dois períodos, de acordo com os procedimentos adotados para a



aquisição de dados: a) de 1860 a 1960, com a utilização de fotografias aéreas; e b) de 1960 até a atualidade, com o uso de satélites e sensores que orbitam a terra (NOVO, 2010). A contemporânea disponibilidade da utilização de imagens de satélite, sobretudo as imagens multiespectrais, viabiliza a aquisição de informações nas distintas partes do espectro eletromagnético, contribuindo, assim, para o reconhecimento e a distinção de objetos de interesse. Dessa forma, o sensoriamento remoto mostra-se como uma valiosa ferramenta para pesquisadores na área de recursos naturais (HENRIQUES, 2008).

Embora as primeiras imagens orbitais datem da década de 1950, seu uso sistemático para fins civis e de pesquisa surgiram a partir de 1972 com o lançamento satélite de tecnologia de recursos terrestres denominado Landsat 1. Seu objetivo primordial foi monitorar os ecossistemas da terra, bem como quantificar as mudanças de sua cobertura (BOYD; DANSON, 2005; XIE *et al.*, 2008; HANSEN *et al.*, 2008). No decorrer das décadas seguintes, vários outros sistemas (plataforma orbital e sensores) com diferentes características foram lançados e seus produtos permitiram também o avanço nas técnicas de transmissão e de processamento dos dados. Entre os vários satélites lançadas neste interim, em 2015, a Agência Espacial Europeia (ESA) lançou o satélite Sentinel 2, com o sensor *Multispectral Imager* (MSI) com 13 bandas espectrais, que oferece uma resolução espacial ainda mais refinada para as bandas do visível e infravermelho próximo (10 metros), cujas imagens são utilizadas no monitoramento da cobertura da terra, detectando suas mudanças no tempo (ESA, 2020).

Além das diferenças nos quesitos resolução espacial e espectral, os produtos de imagens de satélites variam também na cobertura temporal e amplitude de visada – fatores que influenciam a disponibilidade de dados e a escolha do produto. Uma vez combinadas, essas características são decisivas para aplicação na área ambiental (TURNER *et al.*, 2003; HANSEN *et al.*, 2008; GOETZ, 2007; LOARIE *et al.*, 2007).

Outros dados de SR obtidos também por sensores que orbitam a terra são as imagens geradas com a utilização de um sistema de *Radio Detection and Ranging* (RADAR). Trata-se da detecção de ondas de rádio para as imagens do *Shuttle Radar Topography Mission* (SRTM) – missão conjunta da Administração Nacional de Aeronáutica e Espaço (NASA) com a Agência Nacional de Inteligência Geoespacial

(NGA) e as Agências Espaciais Alemã e Italiana para o ano 2000. O objetivo da missão foi produzir um modelo de elevação da superfície terrestre, cobrindo quase todo o planeta. Para tanto, foi utilizada a resolução espectral e o comprimento de onda de ondas da banda C, de 5,6 cm, com resolução espacial de 30 metros e resolução radiométrica inteira de 16 bits (USGS, 2020). Utilizando as imagens SRTM, é possível gerar os Modelos Digitais de Elevação (MDE), que se apresentam como uma matriz com um valor de elevação em cada pixel da imagem ou célula, apresentados como arquivos estruturados em linhas e colunas georreferenciados e que envolvem dados altimétricos (VALERIANO, 2008).

Nesse ínterim, pode-se perceber que o avanço das características dos sensores orbitais e o melhoramento dos algoritmos de classificação vêm aumentando o interesse nos dados de SR para a produção de mapeamentos do uso e cobertura da terra (ADAM *et al.*, 2014). Cumpre lembrar que métodos tradicionais de classificação de imagens, assentados em técnicas estatísticas tradicionais como a máxima verossimilhança e mínima distância, ainda vêm sendo utilizados frequentemente na classificação de imagens de satélites. Contudo, a capacidade limitada desses métodos na resolução de problemas de interclasses tem levado pesquisadores à procura de novos métodos na área da Inteligência Computacional (IC) (MATHER, 1999).

A aprendizagem de máquina (*Machine Learning*) – um subcampo da IC – está relacionada ao aprimoramento de algoritmos e técnicas que auxiliam as máquinas a aprender (ALPAYDIN, 2004). Ao contrário dos tradicionais aplicativos pautados em *pixels*, os classificadores com abordagens baseadas em algoritmos fundamentados em aprendizagem foram desenvolvidos para que fosse possível obter mais informações confiáveis, assim como precisão nas imagens de satélite (AKAR *et al.*, 2012).

Nesse sentido, neste trabalho serão abordadas técnicas de SR que passam pelo processamento digital de imagens (PDI). As técnicas selecionadas envolverão o empilhamento das bandas utilizadas para o satélite Sentinel 2, com 6 bandas espectrais do sensor (MSI), e adicionalmente: a) imagens altimétricas SRTM; b) declividade e rugosidade derivadas do SRTM; c) escolha das amostras; e d)

classificação das imagens empregando três algoritmos de aprendizado de máquina para a classificação do uso e cobertura da terra.

Posteriormente, elas serão classificadas e ordenadas de acordo com a avaliação do desempenho desses algoritmos, através de técnicas matemáticas e estatísticas. Como produto, serão gerados mapas temáticos de uso e cobertura da terra para a área de estudo, que servirão como base de dados para o estudo que compreende a bacia hidrográfica à montante da barragem de Nova Ponte/MG. Dessa monta, será possível apontar áreas-alvo para adequações na gestão e no planejamento da dinâmica da paisagem e dos recursos naturais.

Assim sendo, esta pesquisa parte da premissa de que o empilhamento de imagens de altimetria, como as do SRTM, e derivações, como rugosidade e declividade, integrados em um cubo de dados, propiciarão uma melhoria na classificação de imagens de satélite – neste caso do Sentinel 2 – em relação ao uso apenas na de uma cena.

Para que seja possível atestá-la, o estudo parte da seguinte questão norteadora: dentre os atuais classificadores de imagens que trabalham com aprendizado de máquina selecionados para esta pesquisa, qual terá o melhor desempenho para a área de estudo na classificação, levando-se em conta a utilização das imagens Sentinel-2 e SRTM?

Este estudo justifica-se pela importância de conhecer e quantificar o desempenho dos algoritmos de aprendizado de máquina frente a disponibilidade e diversidade de dados geográficos existentes para o mapeamento de uso e cobertura da terra no Cerrado. A pesquisa busca identificar qual deles tem o melhor desempenho na classificação, observando os resultados da matriz de confusão e dos índices utilizados para a validação, frente a utilização de imagens Sentinel 2 e SRTM.

Para a área de pesquisa, que foi escolhida pela necessidade do levantamento atual, ano 2019, do uso e cobertura da terra para compor a base de dados produzidos na bacia hidrográfica de Nova Ponte/MG, que vem sendo estudada desde o ano de 2009 pelo orientador deste estudo, professor Diego Macedo, com projetos de pesquisa voltados para a avaliação de qualidades ambientais aquáticas e suas relações com o clima e o uso da terra. O conhecimento sobre os algoritmos pode nortear estudantes e pesquisadores da área de SR na escolha dos métodos que

resultem em melhor desempenho, levando-se em conta o tempo gasto de processamento dos dados e os níveis de concordância entre as classes evidenciadas nos índices de validação.

No plano institucional, este estudo configura-se como importante porque pode permitir esta geração de base de dados para o ano de 2019 sobre o uso e cobertura da terra para o projeto de pesquisa *PROECOS - Processos ecológicos - Desenvolvimento de novas ecotecnologias de diagnóstico e processos ambientais*, no qual eu faço parte como pesquisador, que é um projeto que visa produzir modelos ecológicos preditivos sobre as mudanças globais de clima e uso da terra, bem como a relação com o grau de conservação de bacias de reservatórios hidrelétricos. O projeto é financiado pela Agência Nacional de Energia Elétrica (ANEEL) e pela Companhia Energética de Minas Gerais (CEMIG).

Por fim, quanto à relevância social, a pesquisa almeja possibilitar que gestores públicos e setor privado, por meio dos mapas temáticos do uso e cobertura da terra, direcionem a utilização dos recursos naturais da bacia hidrográfica visando a uma melhor gestão e planejamento, assim como incentivos a ações privadas que caminham no rumo das ações de conservação da biodiversidade e avanço do desenvolvimento sustentável.

O produto técnico resultante dos esforços deste trabalho será o mapeamento do uso e cobertura da terra para área de estudo, que compreende a bacia hidrográfica que circunda a barragem de Nova Ponte/MG. Trata-se de uma área de relevante importância ambiental, provedora de serviços ecossistêmicos, e que abriga uma parte do Parque Nacional da Serra da Canastra – uma Unidade de Conservação (UC) – e diversas nascentes e cursos d'água importantes para a macro bacia hidrográfica do Paraná. Os resultados, além de identificar o algoritmo que obteve o melhor desempenho dentro dos índices de validação aplicados na classificação, servirão para demonstrar o retrato e os caminhos tomados até então para o uso e cobertura da terra no ano de 2019.

Ainda, pode-se pontuar que a replicação e a reprojeção da pesquisa para outras áreas será também possível, uma vez que, para a realização deste estudo, foram adotadas metodologias e recursos totalmente gratuitos, como imagens de satélites e softwares de Sistemas de Informações Geográficas (SIG), que podem ser adquiridos

por qualquer pessoa e em qualquer parte do mundo. O software utilizado para o tratamento dos dados e levantamento das classes de uso e cobertura da terra foi o *QGIS 3.16 Hannover*, por ser uma plataforma de geoprocessamento livre e aberta. Outro software empregado neste estudo para a classificação das imagens foi o *Orfeo ToolBox (OTB)*, que também é um projeto de código aberto para Sensoriamento Remoto (SR).

## **1.1 Objetivos**

### **1.1.1 Objetivo Geral**

Analisar e comparar o desempenho de três algoritmos baseados em aprendizagem de máquina - *Random Forest (RF)*, o *Decision Tree (DT)* e o *Support Vector Machine (SVM)* – para a classificação da cobertura do solo no cerrado a partir de um cubo de imagens de SR composto por dados de imageamento, modelo digital de terreno e produtos.

### **1.1.2 Objetivos Específicos**

- Mapear o uso e cobertura da terra na bacia hidrográfica que circunda a barragem de Nova Ponte utilizando duas imagens com cubos de dados distintos, uma empilhando as bandas disponíveis no satélite Sentinel 2 e outra empilhando estas bandas com imagens de altimetria e derivações como declividade e rugosidade.
- Avaliar se o uso de um modelo digital de terreno e variáveis derivadas em conjunto com as imagens do Sentinel 2 determina melhoras nos desempenhos dos resultados das classificações, analisando os índices estatísticos de validação e concordância entre as classes na matriz de confusão.
- Comparar os dados de validação e os mapas das imagens classificadas por meio dos algoritmos abordados com o recorte da área de estudo do mapa

mundial de uso e cobertura da terra para o ano de 2020 produzido pela *Environmental Systems Research Institute (ESRI)*.

## **2 BASES CONCEITUAIS**

### **2.1. Sensoriamento Remoto e Processamento Digital de Imagens no mapeamento do uso e cobertura da terra**

Em uma definição clássica, o SR compreende a arte ou a ciência de adquirir informações de um objeto não existindo contato físico direto com o alvo, mas, sim, e na maioria dos casos, a quilômetros de distância deste (JENSEN, 2009). Nesse sentido, o SR pode ser usado para monitorar e medir características biofísicas e atividades humanas na terra, produzindo imagens como produtos típicos.

A classificação de imagens de sensores orbitais pode ser descrita como uma etapa do método que permite transformar imagens constituídas por valores numéricos em uma carta temática, onde os objetos desta imagem são agrupados e identificados conforme seus aspectos espectrais, espaciais e temporais, transferindo cada *pixel* a uma determinada categoria ou classe anteriormente estabelecida pelo pesquisador (MENEZES; SANO, 2012). Dessa forma, as abordagens utilizadas para a classificação das imagens contidas nos sensores orbitais baseiam-se em metodologias oriundas de processos de: a) caracterização; b) gestão ambiental; c) comparação de paisagens; e d) ordenamento territorial. Todas elas são empregadas usualmente para a modelagem e observação das mudanças nas paisagens (STEVIC *et al.*, 2016).

Para um melhor entendimento dos fatores que causam mudanças na superfície do solo, é essencial que se faça o levantamento e o mapeamento do uso e cobertura do solo por meio de geotecnologias, a exemplo do Sistema de Informações Geográficas (SIG) e do SR. Os incentivos para a execução de projetos geotecnológicos vêm da necessidade de conhecer, mapear e monitorar as migrações internas da população e a exploração dos recursos naturais, que demandam a utilização de tais recursos de forma racional, assim como a ocupação mais ordenada do espaço (COUTINHO, 1997).

O monitoramento dos fenômenos naturais dinâmicos do meio ambiente e o impacto das alterações proporcionadas pelas atividades antrópicas – erosão do solo, inundações, etc. – proporcionam, por meio do SR com imagens orbitais, a reprodução e o registro dessas mudanças nas paisagens, identificando, calculando e monitorando áreas desmatadas, impermeabilizadas ou susceptíveis a processos erosivos (AMARAL; RIOS, 2012).

Com o avanço tecnológico, o registro e a apresentação dos dados de monitoramento e imagens de modo agregado vêm sendo possível, ao que se nomeia a formação de “cubos de dados”, em virtude de sua multidimensionalidade. A seção que se segue evidencia melhor essa técnica e sua relevância.

### **2.1.2 Cubo de dados**

A fusão de imagens de sensores orbitais como um cubo de dados é uma das técnicas utilizadas no processamento digital de imagens (PDI). Ao permitir a agregação de mais informações em uma imagem final única, seu caráter de multidimensionalidade contribui para o progresso das ferramentas de visualização de dados (YOCK, 1996).

Enquanto técnica, os cubos de dados são a conjunção de imagens utilizando dois ou mais sensores, obtendo como resultado uma imagem final melhorada sob diversos aspectos, uma vez que permite agregar uma maior quantidade de dados sobre os diferentes alvos (YOCK, 1996). Atualmente, os processos de empilhamento, co-registro e fusão de dados de produtos ópticos com produtos de radar estão sendo sobrepostos para conseguir melhorar a qualidade da classificação de imagens e de sua precisão (TAVARES *et al.*, 2019).

Outra alternativa para aumentar a precisão dos resultados da classificação final é a integração de índices ópticos que são calculados por meio da razão entre as combinações de diferentes dados espectrais dos alvos. Ao final, são gerados mapas de vegetação, de solo, de água e outros. Os exemplos descritos na literatura incluem: a) o Índice de Vegetação da Diferença Normalizada (NDVI); b) o Índice de Vegetação Ajustado para os Efeitos do Solo (SAVI); e c) o Índice de Umidade da Diferença Normalizada (NDWI) (AKAR; GUGOR, 2015; GHOSH; BEHERA, 2018).

Nesse âmbito, as imagens geradas por radares de abertura sintética (SAR), diferentemente dos sensores ópticos, não passam por muitas interferências decorrentes de variações atmosféricas climáticas como formações de nuvens e de sólidos em suspensão, tornando-se fontes eficazes na identificação dos alvos para a classificação (FORQUOR *et al.*, 2014; DU *et al.*, 2015). Sobre esses aspectos do processo de classificação, estudos como os de Rosenthal *et al.* (1984) e Brisco *et al.* (1989) verificaram um aumento real de 20% e 25%, respectivamente, na precisão geral da classificação quando as imagens ópticas e de radar foram fundidas para o mapeamento de culturas. O uso de cubo de dados tem se caracterizado como uma técnica eficiente no processamento digital de imagens.

### 2.1.3 Classificadores

A aprendizagem de máquinas é uma subárea da Inteligência Computacional que está orientada para o desenvolvimento de algoritmos que permitem às máquinas mapear padrões a partir dos dados e, através desses padrões, registrar regras e tomar decisões. Uma definição apropriada de aprendizado dentro do contexto da Inteligência Computacional é o processo em que um sistema inteligente passa por estímulos do ambiente no qual está inserido, determinando o tipo de aprendizado através da maneira pela qual esta adaptação é realizada (MENDEL; MCLAREN, 1994). O processo de aprendizado envolve a análise dos aspectos como o Sistema Inteligente (SI), cujos resultados dos estímulos recebidos levam ao seu ajuste e respondem, de uma nova maneira, ao ambiente em relação a esses ajustes (HAYKIN, 1999).

Atualmente, os algoritmos de classificação avançados têm alcançado progresso significativo na classificação de imagens. Alguns deles são: rede neural (*Artificial Neural Network*); árvore de decisão (*Decision Tree*); máquina de vetores de suporte (*Supported Vector Machine – SVM*); algoritmos baseados em objetos; algoritmos baseados em sub-*pixel*; e algoritmos contextuais (TSO; MATHER, 2001; LU; WENG, 2007; BLASCHKE, 2010). Ao contrário dos tradicionais aplicativos pautados em *pixels* (AKAR *et al.*, 2012), os classificadores com abordagens baseadas em algoritmos e fundamentados em aprendizagem foram desenvolvidos para obter mais informações confiáveis de imagens de satélite e precisão. Entre os algoritmos



apoiados em aprendizagem, para além dos já mencionados, destacam-se ainda: Floresta Aleatória (*Random Forest*); Ensacamento (*Bagging*); Impulsionamento (*Boosting*); e K-vizinho mais próximo.

Os *métodos de classificação de conjuntos* são algoritmos de aprendizado que constroem um conjunto de classificadores – em vez de apenas um – e, em seguida, classifica novos pontos de dados votando em suas previsões. Entre os classificadores de conjuntos mais usados estão o Ensacamento (*Bagging*), Impulsionamento (*Boosting*) e o Floresta Aleatória (*Random Forest*).

Para o algoritmo *Bagging*, muitas amostras de inicialização (*bootstrap*) são desenhadas a partir de um conjunto de dados de treinamento com substituição para apreender um classificador, ao passo em que uma árvore (estrutura semelhante a uma árvore) é construída para cada amostra de *bootstrap*. Como árvores sucessivas são construídas independentemente das árvores anteriores, um voto majoritário simples é usado para a previsão (LIAW; WINIER, 2002).

Em contrapartida, o *Boosting* utiliza o re-treinamento iterativo e os pesos dos classificados incorretamente, pois as amostras são aumentadas à medida que as interações progridem, para que possam torná-las mais importantes nas próximas interações. Normalmente, o *Boosting* reduz a variação e o viés da classificação, e, em alguns casos, é considerado mais preciso que o *Bagging* (AKAR, 2012).

Tomando-se o montante atual de algoritmos de classificação de imagem baseados em aprendizagem de máquina disponíveis, considerou-se que, para o desenvolvimento desta pesquisa, a comparação de três deles seria o mais adequado. Dessa forma, foram selecionados os seguintes algoritmos: Floresta Aleatória, Árvore de Decisão e Máquina de Vetores de Suporte. A escolha por esses algoritmos foram a relevância e a grande quantidade de estudos realizados em diferentes países utilizando estes classificadores de imagens. A partir daí, a comparação entre os escolhidos deu-se pela eficiência destes sistemas em aprender com os dados, identificar padrões e tomar decisões com pouca intervenção humana (VICERI, 2021).

- **Árvores de Decisão – *Decision Tree (DT)***

Um classificador de Árvores de Decisão caracteriza-se como classificador não paramétrico, não exigindo que sejam feitas suposições estatísticas *a priori* em relação à distribuição de dados. A estrutura básica de uma árvore de decisão consiste em um nó raiz, vários nós internos e um conjunto de nós terminais. De acordo com o quadro de classificação, os dados são divididos recursivamente na árvore de decisão (Figura 1). Em cada nó é necessária uma regra de decisão que pode ser implementada usando um teste de divisão (QUINLAN, 1993).

$$\sum_i^n a_i x_i \leq c \quad \text{para árvores de decisão multivariadas ou simplesmente}$$

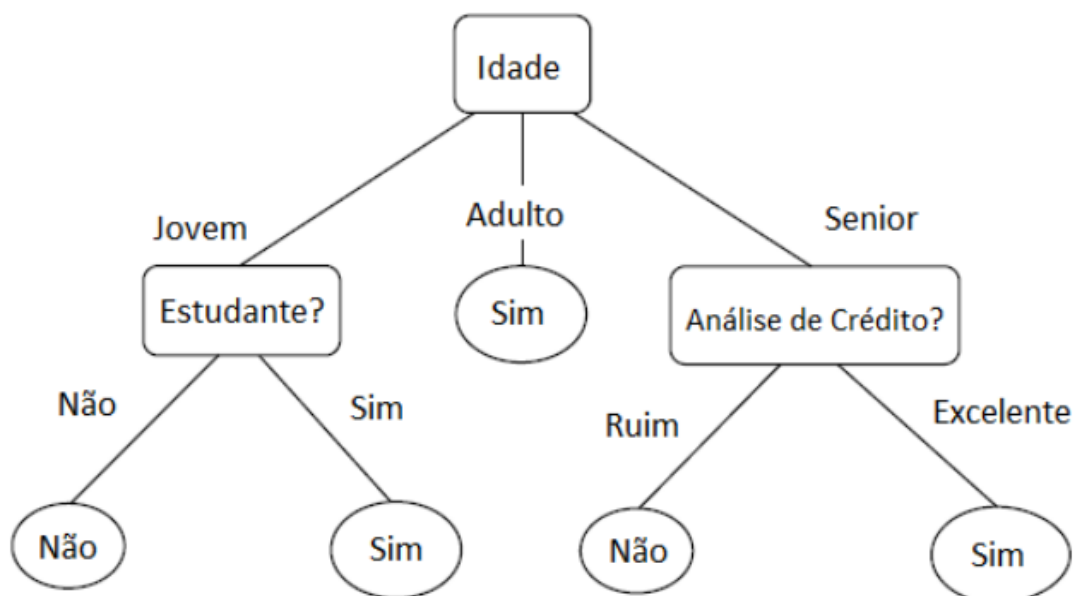
$x_i > c$  para árvores de decisão univariadas.

Onde:

$X_i$  representa os vetores de medição no  $n$  de recursos selecionados;

$\alpha$  é um vetor de coeficientes de discriminação linear;

$c$  é o limiar de decisão (BRODLEY & UTGOFF, 1992).



**Figura 1** - Representação do método de Árvore de Decisão  
Fonte: HAN *et al.* (2011).

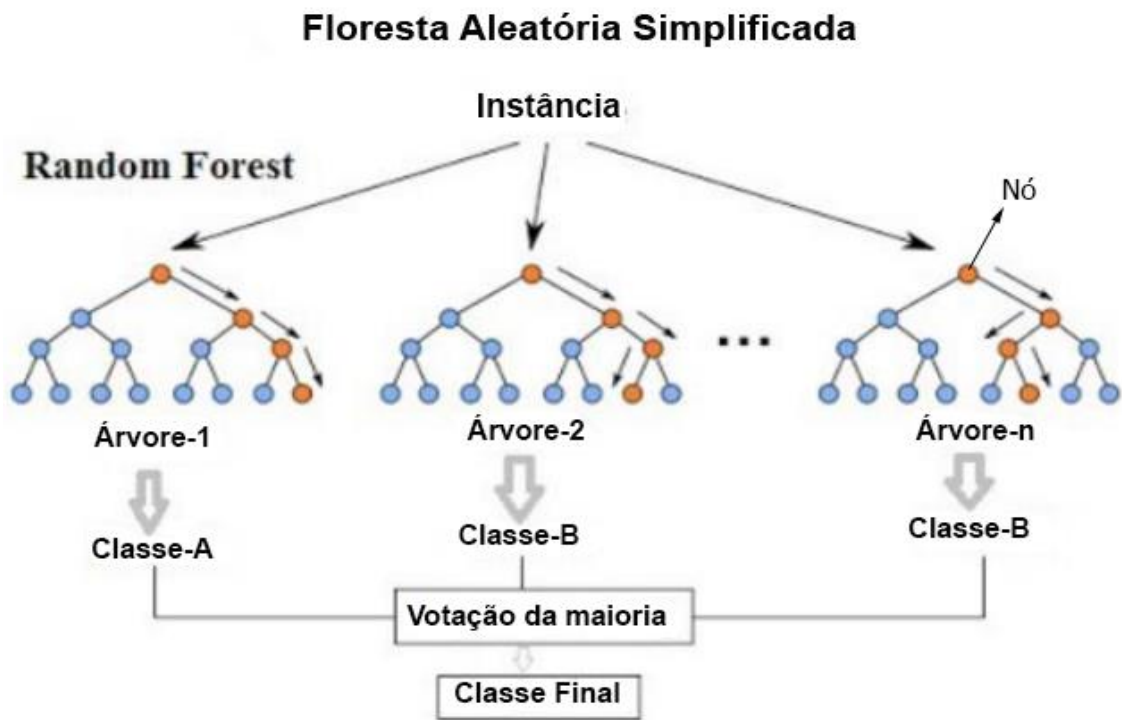
De acordo com Kulkarni *et al.* (2016), os classificadores de árvores de decisão são mais eficientes que os classificadores denominados de estágio único. As decisões

são tomadas em vários níveis, sendo conhecidos também como classificadores de vários níveis. Ao projetar um classificador de árvore de decisão é desejável construir a árvore ideal para alcançar a maior precisão da classificação com o número mínimo de cálculos. As preocupações básicas do classificador de árvore de decisão são as separações de grupos em cada terminal (nó) e as escolhas dos recursos mais eficazes, separando os grupos de classes (KULKARNI *et al.*, 2016).

- **Floresta Aleatória – *Random Forest* (RF)**

Considerado como uma versão mais avançada do *Bagging* (BREIMAN, 2001), o classificador de *Random Forest* (RF) pode ser descrito como classificadores estruturados em árvore, de modo que a aleatoriedade seja adicionada a ele. Metodologicamente, o RF divide cada nó utilizando o melhor dentre um subconjunto de preditores escolhidos aleatoriamente neste nó. Um novo treinamento do conjunto de dados é criado a partir do conjunto de dados original com substituição, cultivando uma árvore a partir da seleção aleatória de recursos (Figura 2).

Suponhamos que  $N$  seja escolhido como 1000. O algoritmo de RF gera 1000 árvores que significam 1000 classificações diferentes de resultados para cada *pixel* específico. Suponhamos novamente que este *pixel* é classificado como floresta em 800 árvores, como terra em 100 árvores e como água em 100 árvores. Por votação majoritária, a saída prevista para o resultado desse *pixel* é floresta (AKAR, 2012).



**Figura 2-** Representação do método de Florestas Aleatórias  
 Fonte: adaptado de MEDIUM (2018).

O usuário deve definir dois parâmetros para inicializar o algoritmo de *Random Forest* (AKAR, 2012). Esses parâmetros são o número de árvores a crescer ( $N$ ) e o número de variáveis usadas para dividir cada nó ( $m$ ). Primeiramente,  $N$  amostras de *bootstrap* são retiradas dos 2/3 do conjunto de dados do treinamento. O 1/3 restante dos dados de treinamento - chamados de dados fora da bolsa (OOB) - são usados para testar o erro das previsões. Dessa forma, uma árvore não podada de cada amostra de auto inicialização é aumentada de forma que em cada nó  $m$  preditores são selecionados aleatoriamente como um subconjunto de preditores variáveis. Por fim, a melhor divisão entre essas variáveis é escolhida (AKAR, 2012).

- **Máquina de Vetores de Suporte – *Support Vector Machine* (SVM)**

As máquinas de vetores de suporte (SVM) são também classificadores não paramétricos. Os SVM são um conjunto de algoritmos de aprendizagem usados para a classificação e regressão. Conforme Otukei (2010), a teoria do SVM foi proposta originalmente por Vapnik e Chervonenkis (1997) e debatida mais tarde com detalhes

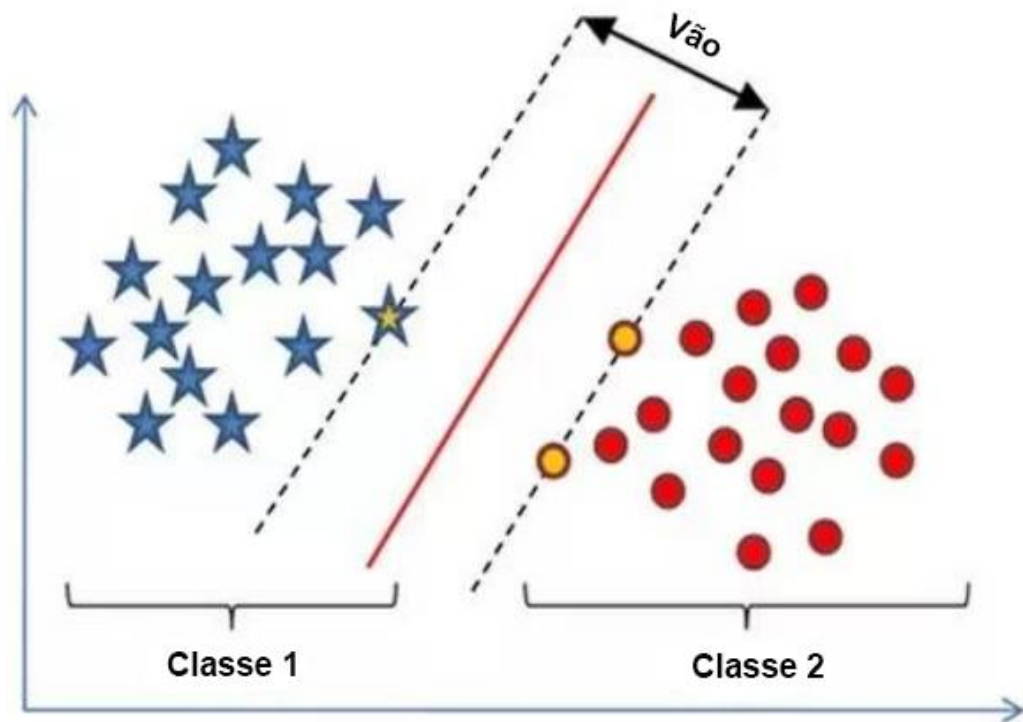
por Vapnik (1999). Para adquirir êxito no SVM, o processo depende diretamente de quão bem é treinado. A forma tida como mais fácil de treinar esse classificador é usar linearmente classes separáveis.

Segundo Osuna *et al.* (1997), se o treinamento de dados com número  $k$  de amostras são representados como  $\{X_i, y_i\}$ ,  $i = 1, \dots, k$  onde  $X \in \mathbb{R}^N$  é um espaço  $N$ -dimensional e  $y \in \{-1, +1\}$  é um rótulo de classe, essas classes são linearmente separáveis se existe um vetor  $W$  perpendicular ao hiperplano linear (que determina a direção do plano discriminante) e um escalar  $b$  mostrando o deslocamento do hiperplano discriminante da origem. Dois hiperplanos podem ser usados para discriminar os pontos de dados nas respectivas classes. Para ambas as classes – i.e. classe 1 representada como  $-1$  e a classe 2 representada como  $+1$  – os hiperplanos serão expressos da seguinte forma:

$$WX_i + b \geq +1 \text{ para todos os } y = +1, \text{ i.e. um membro da classe 1}$$

$$WX_i + b \leq -1 \text{ para todos os } y = -1, \text{ i.e. um membro da classe 2}$$

O objetivo é descobrir em que classe os novos pontos de dados caem. Os dois hiperplanos são selecionados para não apenas maximizar as distâncias entre as duas classes dadas, mas também para não incluir quaisquer pontos entre eles (Figura 3). Em comparação com as abordagens tradicionais, os SVM em geral são relatados como produtores de resultados de maior precisão. Entretanto, isso depende do *kernel* usado, da escolha dos parâmetros para o *kernel* escolhido e o método usado para o SVM gerado (HUANG *et al.*, 2002).



**Figura 3** - Representação do método de Máquina de Vetores de Suporte Fonte: adaptado de QUORA (2018).

Conforme Bonesso (2013), o classificador SVM é fundamentado no princípio da minimização do Risco Estrutural (*Structural Risk Minimization – SRM*) proposto por Vapnik e Chervonenkis (1974) e toda sua equipe. A definição do SRM seria a função  $f(x)$  que minimiza o risco médio no conjunto de treinamento, buscando minimizar: a) o erro com relação a um conjunto de treinamento (risco empírico); e b) o erro com relação ao conjunto de testes amostras não empregadas no treinamento do classificador (risco de generalização).

O objetivo do SVM está em obter um equilíbrio entre esses erros, minimizando o excesso de ajustes (*overfitting*) que podem reduzir a capacidade de generalização do classificador. Em vez de extrair características gerais que permitirão a generalização ou reconhecimento de padrões não utilizados no treinamento do classificador, o problema do *overfitting* está relacionado ao fato de o classificador memorizar os padrões de treinamento, gravando suas peculiaridades e ruídos (SMOLA *et al.*, 2000).

## 2.2 Validação da classificação e coeficientes estatísticos

As validações das classificações serão desenvolvidas por meio de dados estatísticos como os índices *kappa*, exatidão global e erro de omissão e comissão, utilizando os dados da matriz de confusão gerada na classificação das imagens e as amostras que vão servir para realizar o teste da classificação.

A matriz de confusão oferece a base de dados para representar a precisão da classificação e definir os erros. Dentro de uma matriz são derivadas algumas medidas de correção para a classificação, como a exatidão global (FOODY, 2002). Comparada com a categoria verificada em campo e o número de unidades de amostras de uma categoria particular relativa compreendido por um classificador (ou regra de decisão), a matriz de confusão é construída por um esboço quadrado de números dispostos em linhas e colunas (CONGALTON, 1991). Na matriz de confusão, o nível de acerto é indicado pelos elementos da diagonal principal, demonstrando a concordância entre esses dois conjuntos de dados (Tabela 1).

**Tabela 1:** Caracterização matemática da matriz de confusão

i=linhas (classificação)	j=colunas (referência)			Total das linhas $n_{i+}$
	1	2	k	
1	$n_{11}$	$n_{12}$	$n_{1k}$	$n_{1+}$
2	$n_{21}$	$n_{22}$	$n_{2k}$	$n_{2+}$
k	$n_{31}$	$n_{32}$	$n_{3k}$	$n_{k+}$
Total das colunas $n_{+j}$	$n_{+1}$	$n_{+2}$	$n_{+k}$	N

Fonte: Adaptado de Bernardes (2006).

Onde:  $n$  amostras estão divididas em  $k^2$  células, que no caso são os pixels, sendo que cada amostra está associada a uma das  $k$  classes de classificação, geralmente como linhas no gráfico, para uma das  $k$  classes dentro do conjunto de dados como referência, geralmente como colunas no gráfico. O número de amostras representadas como  $n_{ij}$  classificadas na classe  $i$  ( $i=1, 2, \dots, k$ ) na classificação e classe  $j$  ( $i=1, 2, \dots, k$ ), como o conjunto de dados de referência (BERNARDES, 2006).

$$\text{Sendo } n_{i+} = \sum_{j=1}^k n_{ij}$$

o número de amostras classificadas dentro da classe  $i$  na simulação de classificação,

$$\text{e } n_{+j} = \sum_{i=1}^k n_{ij}$$

o número de amostras classificadas dentro da classe  $j$  no conjunto de referência.

O índice *kappa* é um método estatístico para avaliar o nível de concordância ou reprodutibilidade entre dois conjuntos de dados. Para Cohen (1960), a estatística *Kappa* é uma medida de concordância corrigida, subtraindo a proporção observada de concordância pelo nível de concordância esperada de acordo com expressão que segue:

$$K = \frac{n \sum_{i=1}^k n_{ii} - \sum_{i=1}^k n_{i+} n_{+i}}{n^2 - \sum_{i=1}^k n_{i+} n_{+i}}$$

Onde:

$K$  = número de linhas da matriz de erros;

$n$  = número total de observações (amostras);

$n_{ii}$  = número de observações na linha  $i$  e coluna  $i$ ;

$n_{i+}$  = total da linha  $i$ ;

$n_{+i}$  = total da coluna  $i$ .

No caso do índice *kappa*, quanto mais próximo do valor 1, conforme desempenho estabelecido por Congalton e Green (1999) (Tabela 2), maior será a concordância entre os dois conjuntos de dados.

**Tabela 2** - Avaliação do grau de concordância do índice *kappa*



Valor do <i>Kappa</i>	Concordância
$K \leq 0$	Péssimo
$0,0 < K \leq 0,2$	Ruim
$0,2 < K \leq 0,4$	Razoável
$0,4 < K \leq 0,6$	Bom
$0,6 < K \leq 0,8$	Muito Bom
$0,8 < K < 1,0$	Excelente

Fonte: Congalton e Green (1999).

Com os dados contidos na matriz de confusão das imagens classificadas, foi calculada a exatidão global. De acordo com Chuviecco (1995), a exatidão global correlaciona os elementos da diagonal principal com o total de pontos amostrados. Com valores mais altos que os do índice *kappa*, a exatidão global é supervalorizada por não considerar os erros de omissão e comissão que estão contidos na matriz de confusão (BRITES, 1996).

$$F_m = \frac{\sum x_{ii}}{\sum \sum x_{ij}}$$

Onde:

$F_m$  = exatidão global;

$x_{ii}$  = elementos da diagonal;

$\sum \sum x_{ij}$  = soma dos elementos.

Para estimar os erros de omissão (falso negativo) e comissão (falso positivo) também são utilizados os dados inclusos na matriz de confusão. Para tanto, é necessário considerar as células marginais nas linhas, indicando o número de *pixels* que não foram incluídos em determinada categoria, revelando os erros conhecidos como omissão. Conforme Ferreira *et al.* (2007), o significado para a omissão refere-se como uma definição imperfeita da categoria, enquanto a comissão como uma delimitação excessiva da categoria. Os erros de omissão ( $EO_i$ ) e comissão ( $EC_j$ ) são calculados para cada classe temática na classificação de acordo com as expressões que seguem:

$$EO_i = \frac{\sum x_{ij} - x_{ii}}{\sum x_{ij}} \quad EC_j = \frac{\sum x_{ij} - x_{jj}}{\sum x_{ij}}$$

Onde:

$\Sigma x_{ij} - x_{ii}$  = soma dos resíduos por linha;

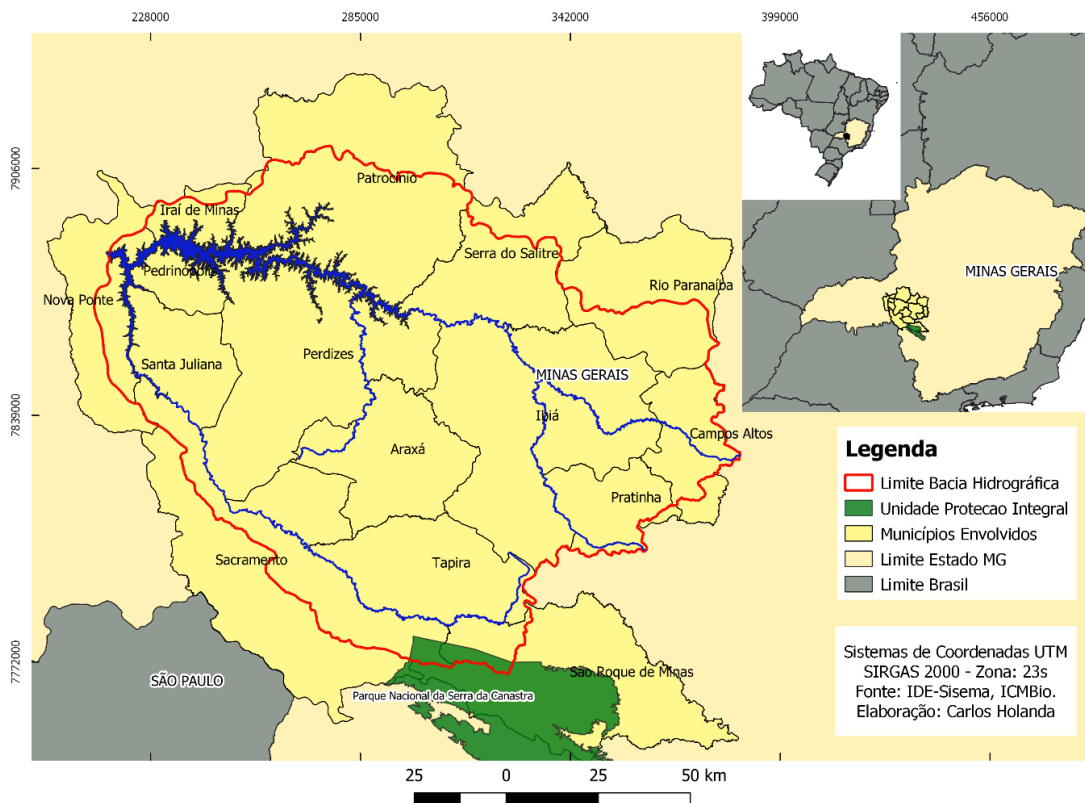
$\Sigma x_{ij} - x_{jj}$  = soma dos resíduos por coluna;

$\Sigma x_{ij}$  = marginal da linha ou coluna.

### **3 CARACTERIZAÇÃO DA ÁREA DE ESTUDO**

A caracterização da área de estudo é de fundamental importância para o entendimento global da configuração histórica, econômica, social e ambiental do espaço. Possibilita compreender não apenas a configuração atual dos diversos usos e coberturas da terra, mas também suas relações com a história e economia, bem como abrir espaço para se projetar novas medidas de gestão e uso da área. Em sentido oposto, serve ainda para se fazer um levantamento socioeconômico e ambiental da região, auxiliando também na elaboração das classes temáticas que irão compor a classificação das imagens de satélite.

A área de estudo desta pesquisa se localiza no bioma Cerrado e compreende o limite da bacia hidrográfica da barragem de Nova Ponte. A extensão total é de 15.358 km<sup>2</sup> e abrange quinze municípios na região do Alto Paranaíba no Estado de Minas Gerais (Figura 4).



**Figura 4** - Mapa de localização da área de estudo  
Fonte: IDE-Sisema, ICMBio.

Tratando de urbanização, situação corriqueira no Brasil também se processa na região: apesar dos dados da população urbana serem maiores que os da população rural, suas economias estão baseadas no agronegócio. A caracterização socioeconômica e ambiental dos municípios que compõem a área da bacia hidrográfica (Tabela 3) mostra que suas economias estão voltadas em maior parte para o cultivo de grãos e a pecuária (IBGE, 2020). Dentre as técnicas de plantio para as diferentes culturas, ganham destaque os pivôs centrais. Alguns municípios, entretanto, têm suas bases econômicas firmadas na mineração e em indústrias de fertilizantes, como é o caso da cidade de Araxá. As cidades que têm os valores de Produto Interno Bruto (PIB) mais elevados são Nova Ponte, Araxá, Sacramento, Santa Juliana, Rio Paranaíba e o município de Tapira – este último com o maior PIB da região (IBGE, 2020).

**Tabela 3** - Dados dos municípios que englobam a bacia hidrográfica (IBGE, 2020)

<b>Municípios</b>	<b><sup>1</sup>Área/ Km<sup>2</sup></b>	<b><sup>1</sup>Pop. Urbana/ Pessoas</b>	<b><sup>1</sup>Pop. Rural/ Pessoas</b>	<b><sup>1</sup>PIB Per capita R\$ anual</b>	<b><sup>1</sup>Principais Atividades Econômicas</b>	<b><sup>2</sup>Clima e Temperatura média anual</b>	<b><sup>3</sup>Relevo Predominante</b>
<b>Nova Ponte</b>	1.111, 011	10.991	1.821	51.809,94	Produção Florestal, Pesca e Agricultura	Média 21,9°C/ Tropical de altitude (Cwb)	Plano a ondulado
<b>Pedrinópolis</b>	357,891	2.927	563	29.115,40	Agricultura e Pecuária	Média 20,0°C/ Tropical de altitude (Cwb)	Plano a ondulado
<b>Patrocínio</b>	2.874, 344	72.758	9.713	30.089,34	Agricultura e Pecuária	Média 20,7°C/ Tropical de altitude (Cwb)	Suave-ondulado a forte-ondulado
<b>Perdizes</b>	2.451, 112	9.935	4.469	37.778,59	Agricultura e Pecuária	Média 19,8°C/ Tropical de altitude (Cwb)	Plano a forte- ondulado
<b>Araxá</b>	1.164, 062	92.284	1.388	49.299,60	Mineração, Fertilizantes, Agropecuária	Média 19,2°C/ Tropical de altitude (Cwb)	Suave-ondulado a forte-ondulado
<b>Ibiá</b>	2.704, 132	19.646	3.572	36.664,73	Agricultura e Pecuária	Média 19,3°C/ Tropical de altitude (Cwb)	Suave-ondulado a forte-ondulado

<b>Sacramento</b>	3.073,268	19.275	4.621	59.030,98	Agricultura	Média 19,3°C/ Tropical de altitude (Cwb)	Suave-ondulado a forte-ondulado
<b>Santa Juliana</b>	723,784	9.795	1.542	47.871,43	Agricultura	Média 19,7°C/ Tropical de altitude (Cwb)	Plano a ondulado
<b>Iraí de Minas</b>	356,264	5.158	1.309	30.637,59	Agricultura e Pecuária	Média 20,3°C/ Tropical úmido (Cwa)	Plano a forte-ondulado
<b>São Roque de Minas</b>	2.098,867	4.222	2.464	27.366,86	Agricultura e Pecuária	Média 18,7°C/ Tropical de altitude (Cwb)	Ondulado a montanhoso
<b>Pratinha</b>	622,478	1.759	1.506	26.429,45	Agricultura e Pecuária	Média 18,7°C/ Tropical de altitude (Cwb)	Ondulado a forte-ondulado
<b>Tapira</b>	1.179,248	2.744	1.368	146.514,59	Indústrias extrativas	Média 18,4°C/ Tropical de altitude (Cwb)	Ondulado a montanhoso
<b>Campos Altos</b>	710,645	12.913	1.293	21.147,86	Agricultura	Média 18,8°C/ Tropical de altitude (Cwb)	Plano a forte-ondulado
<b>Rio Paranaíba</b>	1.352,353	7.288	4.597	40.105,28	Agricultura	Média 19,3°C/ Tropical de altitude (Cwb)	Plano a forte-ondulado

<b>Serra do Salitre</b>	1.295,272	7.755	2.794	30.096,24	Agricultura	Média 19,7°C/ Tropical de altitude (Cwb)	Plano a forte-ondulado
-------------------------	-----------	-------	-------	-----------	-------------	---	------------------------

Fonte: <sup>1</sup>IBGE (2010); <sup>2</sup>Alvares *et. al.* (2013); <sup>3</sup>SRTM (USGS, 2020).

Geograficamente, a área abrange desde regiões alagadas ao norte-nordeste pela barragem de Nova Ponte com altitudes médias de 800 metros – região próxima à cidade de Nova Ponte –, até áreas com altitudes mais elevadas ao sul da bacia hidrográfica, chegando a 1.200 metros de altitude.

O Parque Nacional da Serra da Canastra, unidade de proteção integral que está localizado no extremo sul da bacia hidrográfica, está situado no município de São Roque de Minas, onde, juntamente com Sacramento, são encontradas as maiores altitudes da bacia. A nascente histórica do Rio São Francisco (IBGE, 2020) compõe a mesma paisagem. A antiga cidade de Nova Ponte foi inundada em meados de 1993/1994, criando a nova cidade na parte mais alta da região, hoje ao lado da represa da Usina Hidrelétrica ([novaponte.mg.gov.br](http://novaponte.mg.gov.br)).

### **3.1 Cobertura Vegetal**

Variando muito em sua forma paisagística, o Cerrado engloba desde vegetação rasteira densa, normalmente com uma cobertura esparsa de arbustos e árvores pequenas, até florestas quase fechadas com altura de dossel de 12 a 15 metros. Com a forma característica de savana, as árvores apresentam tipicamente troncos contorcidos, casca grossa e folhas esclerófilas duras, algumas das quais são sempre verdes, enquanto outras são decíduas por períodos variáveis da estação seca (RATTER *et al.*, 1997). O Cerrado possui relações de um complexo vegetacional com características ecológicas e fisionômicas comparável com outras savanas da América, África, Sudeste da Ásia e Austrália (EITEN, 1979). O bioma conta ainda com abundância de florestas de galeria, onde ocorrem solos mais ricos em nutrientes, diferentemente da maior parte dos solos do Cerrado, considerados pobres (RATTER *et al.*, 1997).

A definição de savana elaborada por Cole (1986) compreende uma vegetação com um estrato graminoso contínuo, exibindo características estruturais e funcionais similares, usualmente com árvores e/ou arbustos, com composição florística variável, diferenciando-a fisionomicamente de um campo puro ou arvoredos decíduos. Outros autores definem o Cerrado como planície de regiões quentes com estação seca mais ou menos longa, coberta com uma vegetação baixa de gramíneas, arbustos e árvores

esparsas (HUBER; RIINAR, 1997). A Savana brasileira (Cerrado) é definida como uma vegetação xeromorfa, podendo também ser encontrada em clima ombrófilo, mas existindo um predomínio de clima estacional (por volta de 6 meses secos) (IBGE, 2013). Por essa classificação é dividida em quatro subgrupos de formação:

- Savana Florestada (Cerradão)

Ocorre em clima tropical eminentemente estacional e com solos profundos, restritos das áreas areníticas lixiviadas, com fisionomia típica e característica na sua formação. Apresenta sinúsias lenhosas de micro e nanofanerófilos tortuosos com ramificação irregular, provindas de semidecíduos ou de macrófilos esclerófilos perenes, com órgãos de reserva subterrâneos ou xilopódio, ritidoma esfoliado corticoso rígido ou córtex maciamente suberoso. De norte a sul, a sua florística reflete uma fisionomia caracterizada por dominantes fanerófitos típicos extremamente repetitivos (IBGE, 2013).

- Savana Arborizada (Campo Cerrado, Cerrado Típico)

Apresenta uma característica fisionômica nanofanerofítica rala e outra hemicriptofítica graminóide, contínua e sujeita ao fogo anual, de formação natural e/ou antrópica. Formando uma fisionomia em terrenos degradados, apesar de semelhante a Savana Florestada e de acordo com o espaço geográfico ocupado, sua composição florística possui ecótipos dominantes que caracterizam os ambientes (IBGE, 2013).

- Savana Parque (Campo Sujo e Campo Rupestre)

Integrada por hemicriptófilos e geófitos de florística natural e/ou antropizada, intercalado por nanofanerófitos isolados, essencialmente constituído por estrato graminóide, através de conotação típica de um parque inglês. Natural e de natureza antrópica, é encontrada em todo o país. Possui os seguintes ecótipos: Ilha de Marajó, Pantanal-Sul-Mato-Grossense e Depressão do Araguaia e Ilha do Bananal (IBGE, 2013).

- Savana Gramíneo-Lenhosa (Campo Limpo)



Quando natural, prevalecem nesta fisionomia plantas lenhosas raquílicas com gramados entremeados, utilizando extensas áreas dominadas por hemcriptófilas. Aos poucos, por meio do fogo ou pastoreio, quando manejadas, são substituídas por geófitos que se distinguem apresentando colmos subterrâneos, mais resistentes ao pisoteio do gado e das queimadas. Muito diversificada em sua composição florística, seus ecótipos mais característicos são as plantas lenhosas (IBGE, 2013).

### **3.2 Macrodrenagem da área de estudo**

O principal rio da bacia hidrográfica que compreende a área de estudo é o Araguari, que tem sua nascente na cidade de São Roque de Minas, dentro do Parque Nacional de Serra da Canastra, chegando a uma altitude de 1.327m; e sua foz no Rio Paranaíba, na divisa dos estados de Minas Gerais e Goiás com uma altitude de 506m (CETEC, 1983). Seus principais afluentes são o Rio Uberabinha, Ribeirão das Furnas, Rio Claro, Rio Quebra Anzol, Rio Galheiro, Rio Misericórdia e Rio do Inferno.

O relevo acidentado e o forte potencial energético por causa do seu grande volume de água propiciam a possível construção de usinas hidrelétricas como as de Nova Ponte, Capim Branco I e II, Miranda e Macacos (PMSB, 2014).

## **4 METODOLOGIA**

As metodologias que serão abordadas para a elaboração deste estudo levam em conta técnicas de SR com arquivos de imagens orbitais como Sentinel 2, SRTM e imagens de declividade e rugosidade derivadas da SRTM.

Esses dados complementares serão adicionados aos algoritmos de classificação a serem aplicados. Sua importância para a classificação reside em não contar apenas com os dados oferecidos pelo satélite multiespectral, mas permitir o uso, também, de informações relevantes sobre altimetria, declividade e rugosidade aos algoritmos abordados.

Serão utilizadas técnicas de PDI para o tratamento dessas imagens como: a) equalização dos histogramas; b) mosaicos das cenas para a área de estudo; c) empilhamentos das bandas utilizadas para cada sensor; d) seleção das amostras para

a classificação por meio de imagens do *Google Earth*; e) análise das imagens em diferentes composições de bandas; e f) seleção aleatória das amostras utilizadas para o teste.

A classificação das imagens se dará nos diferentes algoritmos de aprendizado de máquina objetos de estudo da presente pesquisa: *Random Forest* (RF); *Decision Tree* (DT); e *Support Vector Machine* (SVM). A pós-classificação dessas imagens classificadas será realizada para corrigir possíveis dados de *pixel* das imagens que não se enquadraram em algum grupo de classe pré-definida.

Tais técnicas irão permitir posteriores análises reaproveitando-se os dados dessas classificações para avaliação do desempenho desses algoritmos, por meio de métodos matemáticos e estatísticos baseados na matriz de confusão para gerar os índices *kappa*, erro de omissão e comissão e a exatidão global.

Como produto técnico, serão gerados mapas temáticos de uso e cobertura da terra para o ano de 2019 e um banco de dados para possíveis estudos posteriores na área da bacia hidrográfica.

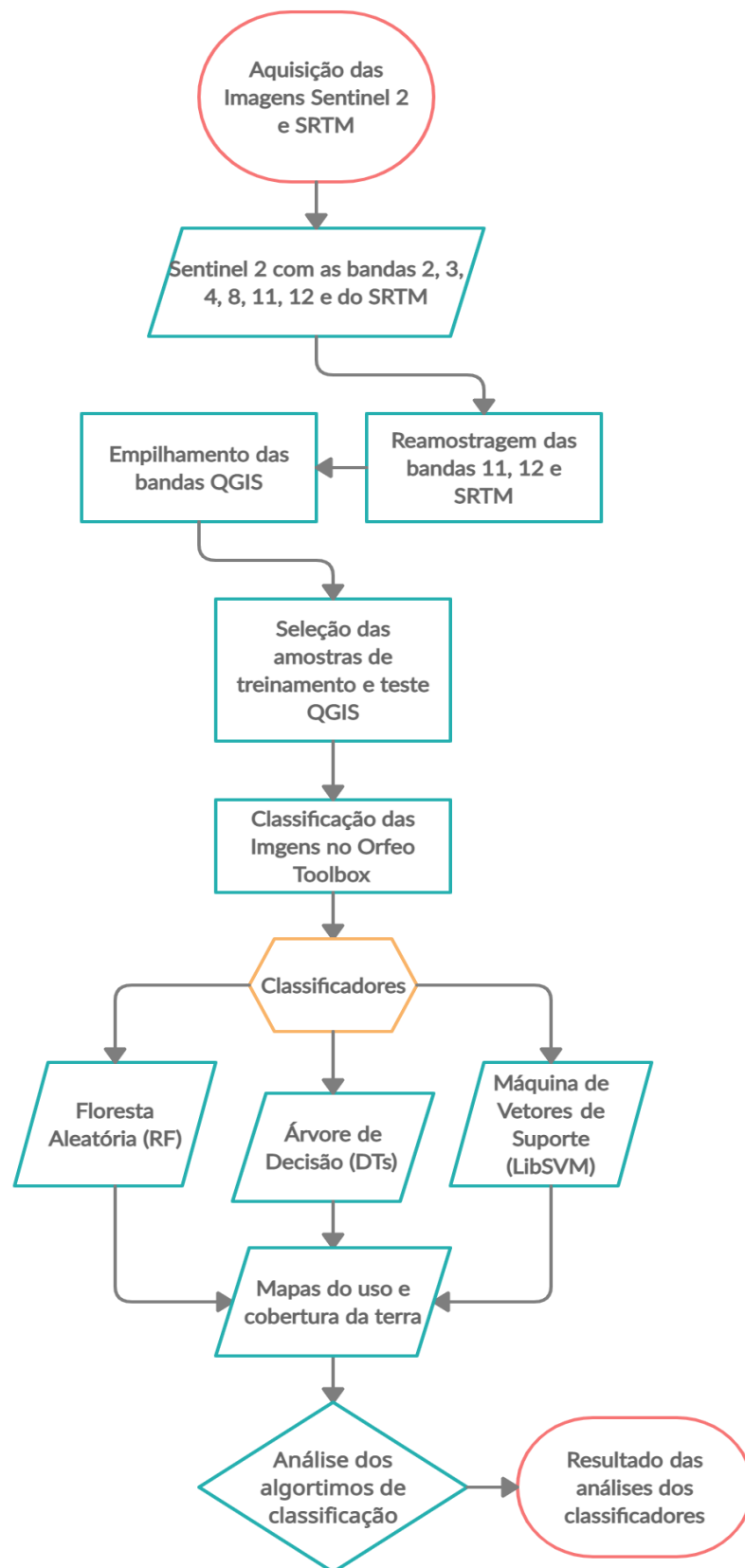
#### **4.1 Recursos materiais**

Serão analisadas, empilhadas, as imagens do satélite Sentinel 2 com uma imagem de altimetria, declividade e rugosidade derivada da missão *Shuttle Radar Topography Mission* (SRTM), obtendo como resultado final a imagem como um cubo de dados. Na análise, serão utilizadas as informações das classificações das imagens por meio do uso e cobertura da terra para produzir os dados para a matriz de confusão. Esses dados servirão para calcular o índice *kappa*, os erros de omissão e comissão e a exatidão global, os quais permitirão identificar qual algoritmo de classificação obteve os melhores desempenhos para as distintas composições. O fluxograma constante da Figura 6 esquematiza o processo.

A análise do uso e cobertura da terra será realizada por meio de imagens de satélites multiespectrais empregando técnicas de SR, PDI e análises sintéticas feitas nas imagens do *Google Earth* do ano de 2019, mesmo ano das imagens de satélite, para fins de identificação e confirmação das áreas e suas respectivas classes. A escolha das imagens para ano de 2019 foi determinada pelo marco das coletas de

dados de qualidade de água obtidos em estudos anteriores para nível de futuras comparações. O mês de julho foi escolhido por ser inverno no hemisfério sul, disponibilizando imagens de satélites com menos quantidades de nuvens e ruídos atmosféricos.

Importante salientar que para a realização deste estudo e o processamento dos dados para a classificação e pós-classificação das imagens de satélite foi utilizado o meu notebook que tem as seguintes características: a) Processador Intel(R) Core(TM) i5-2410M CPU @ 2.30GHz 2.30 GHz; b) Memória instalada (RAM) de 8.00 GB; c) Tipo de sistema, Sistema Operacional de 64 Bits e d) Windows 7 Professional.



**Figura 5** - Fluxograma de atividades desenvolvidas no estudo  
Fonte: Autor.

## 4.2 Pré-processamento das imagens de Declividade e Rugosidade

A utilização das imagens de declividade e rugosidade são importantes para o processo de classificação porque agregam informações ao conjunto de dados disponibilizados pelas bandas do satélite Sentinel 2, auxiliando ainda mais os algoritmos com dados complementares referentes à distribuição das classes. A metodologia usada para gerar a imagem de declividade foi descrita por Horn (1981). Ela emprega uma equação de diferenças finitas de terceira ordem, adequada para gerar uma estimativa de inclinação média dentro de uma grade regular de 3x3 levando em conta o conjunto de oito *pixels* vizinhos em relação ao *pixel* central (KICKEY, 2000). A fórmula para o sistema de informações geográficas (SIG) QGIS é:

- [-p usa inclinação percentual (padrão = graus)] [-s escala \* (padrão = 1)]
- [-alg Horn]
- [-calcular\_bordas] [-b Banda (padrão= 1)] [-de formato] [-co "NOME=VALOR"]\* [-q]

Onde:

*p*: se especificado, a inclinação será expressa como inclinação percentual, caso contrário é expresso em graus.

*s*: proporção de unidades verticais para horizontal.

*calcular\_bordas*: fazer o cálculo nas bordas raster e perto dos valores sem dados.

*b <band>*: selecionando uma banda de entrada a ser processada. As bandas são numeradas a partir de 1.

*de <formato>*: selecionando o formato de saída.

*co "NOME=VALOR"*: muitos formatos têm uma ou mais opções de criação opcionais que podem ser usadas para controlar detalhes sobre o arquivo criado.

*q*: suprime o monitor de progresso e outras saídas sem erros.

Para a metodologia empregada para gerar a imagem de rugosidade, foi utilizado o índice de rugosidade do terreno (TRI) proposto por Riley *et al.* (1999), que é a diferença entre o valor de uma célula da grade e a média das 8 células vizinhas da grade. Com valores calculados a partir de dados de elevação, o TRI gera um *raster* de banda única (WILSON *et al.*, 2007) utilizando a seguinte fórmula:

➤ [-calcular\_bordas] [-b Banda (padrão = 1)] [-de formato] [-q]

Onde:

*calcular\_bordas*: fazer o cálculo nas bordas raster e perto dos valores sem dados.

*b<band>*: selecionando uma banda de entrada a ser processada. As bandas são numeradas a partir de 1.

*de<formato>*: selecionando o formato de saída.

*q*: suprime o monitor de progresso e outras saídas sem erros.

### 4.3 Aquisição de imagens

As imagens obtidas dos dados de elevação (SRTM) foram produzidas por um sistema de *Radio Detection and Ranging* (RADAR), que é a detecção de ondas de rádio pela missão conjunta da Administração Nacional de Aeronáutica e Espaço (NASA) com a Agência Nacional de Inteligência Geoespacial (NGA) para o ano 2000<sup>1</sup>. A partir da imagem SRTM foram geradas as imagens de declividade e de rugosidade utilizando o *software QGIS 3.16 Hannover*, que é uma plataforma de geoprocessamento livre e aberta<sup>2</sup>. As imagens foram empilhadas à imagem do satélite Sentinel 2 e suas respectivas bandas estão mencionadas a seguir:

**Tabela 4** - Dados sobre o satélite Sentinel 2 e o sensor utilizado

Satélite	Sentinel 2

<sup>1</sup> Acessível via <https://earthexplorer.usgs.gov/>

<sup>2</sup> Acessível via [https://www.qgis.org/pt\\_BR/site/](https://www.qgis.org/pt_BR/site/)

Ano de Lançamento	2015 – (ESA)
Sensor	Instrumento Multiespectral (MSI)
Resolução Espacial	10 metros
Resolução Radiométrica	12 bits
Resolução Temporal	5 dias de revisada
Resolução Espectral: Bandas utilizadas e suas respectivas frequências no espectro eletromagnético.	Azul (B2) – 490 a 560 nm, Verde (B3) – 560 a 665 nm, Vermelho (B4) – 665 a 842 nm, Infravermelho próximo (B8) – 842 a 705, Infravermelho médio 1 (B11) – 1610 a 2190 nm e Infravermelho médio 2 (B12) 2190 nm acima.

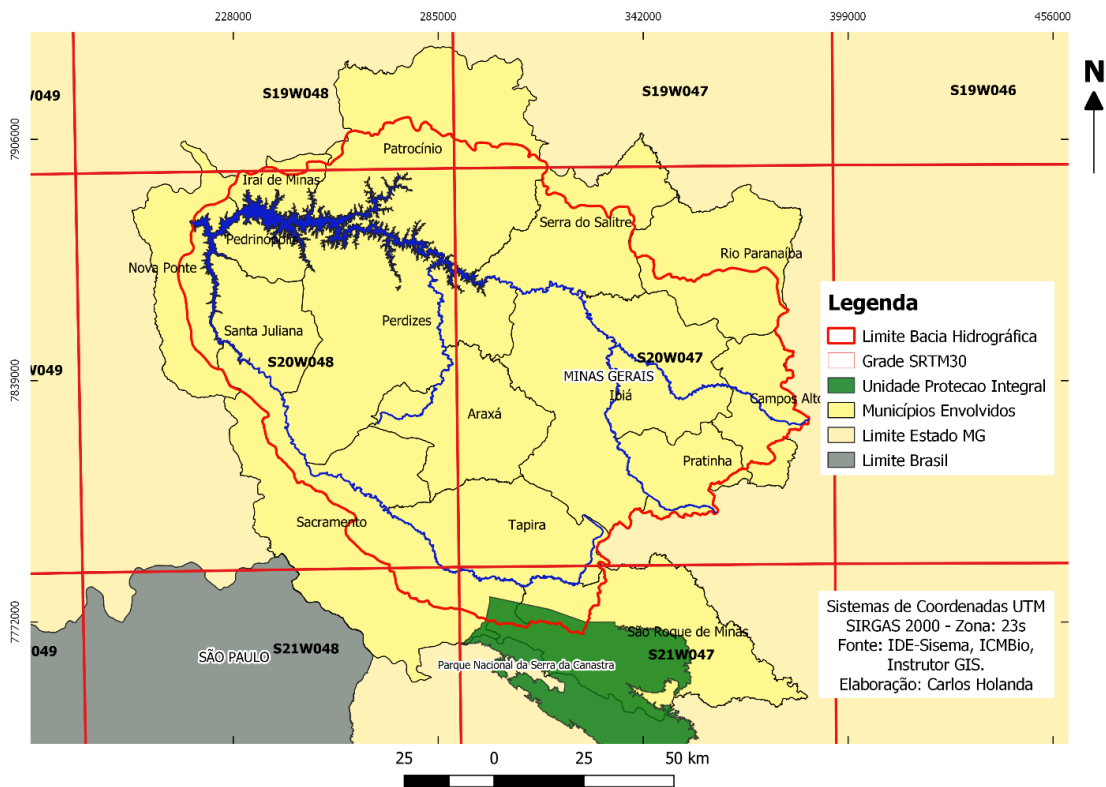
Fonte: Agência Espacial Europeia (ESA, 2020).

Para a aquisição da imagem de radar SRTM, foram realizados *downloads* no sítio do Serviço Geológico dos Estados Unidos (USGS), na interface *Earth Explorer*<sup>3</sup>. Para as imagens do Sentinel 2 foram realizados *downloads* no sítio da página da Agência Espacial Europeia (ESA), dentro da interface Centro de Acesso Aberto *Copernicus*<sup>4</sup>. Ambos os sítios disponibilizam gratuitamente as cenas. Todas as cenas que envolviam a área de estudo do Sentinel 2 e SRTM foram adquiridas com pré-processamento de calibração radiométrica da imagem, remoção de ruído e correção geométrica que calcula os níveis de cinza por interpolação.

Apesar de as imagens do SRTM serem contínuas e geradas por radares de abertura sintética (SAR), foram necessárias seis porções para compor o mosaico da área total para a bacia hidrográfica. As imagens de RADAR (Figura 6) para o ano 2000 incluíram as cenas S19W048, S19W047, S20W048, S20W047, S21W048 e S21W047, todas do dia 11 do mês de fevereiro.

<sup>3</sup> Acessível via <https://earthexplorer.usgs.gov/>.

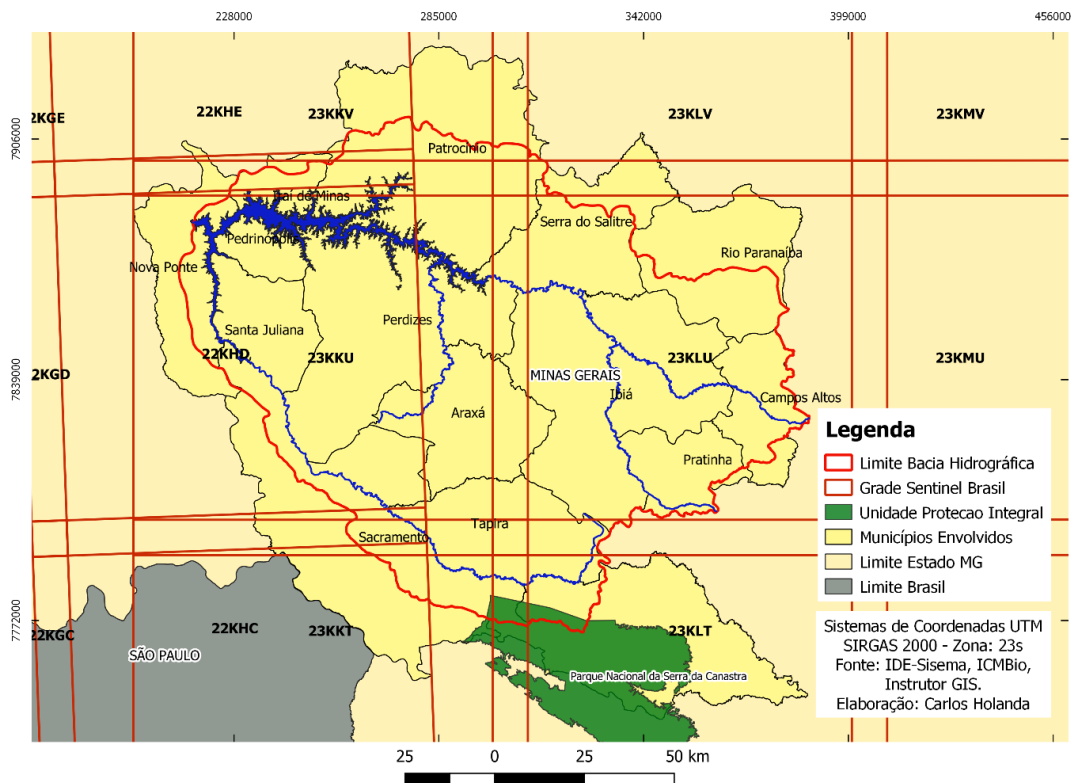
<sup>4</sup> Acessível via <https://scihub.copernicus.eu/dhus/#/home>.



**Figura 6** - Grade de distribuição dos dados do SRTM  
 Fonte: IDE-Sisema, ICMBio e Instrutor GIS.

Para as imagens do Sentinel 2 (Figura 7), foram necessárias seis cenas do ano de 2019 para compor o mosaico da área de abrangência: 23KKV, 23KKU, 23KKT, 23KLV, 23KLU do dia 14 do mês de julho; e 23KLT do dia 04 do mês de julho. Foram empregados o empilhamento das bandas, o mosaico das cenas para todas as imagens, e a reamostragem das imagens de declividade e rugosidade para 10 metros.

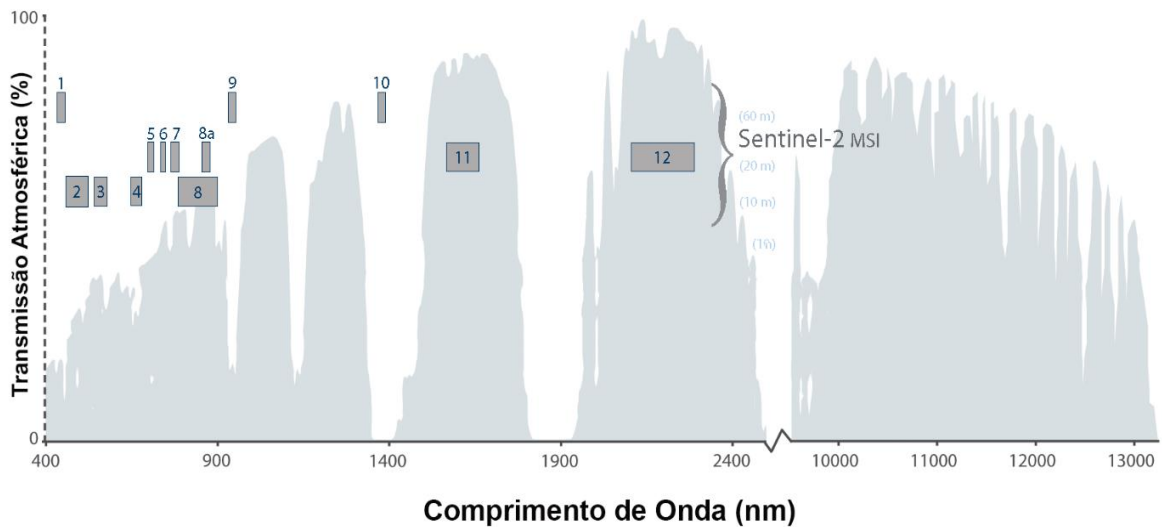




**Figura 7** - Grade das cenas do Sentinel 2  
 Fonte: IDE-Sisema, ICMBio e Instrutor GIS.

Estas imagens do Sentinel 2, empilhadas também com as imagens *SRTM*, declividade e rugosidade foram utilizadas: a) as bandas 2, 3 e 4 correspondentes às bandas do azul, verde e vermelho do visível; b) a banda do infravermelho próximo (*NIR*), que equivale a banda 8; e c) as bandas 11 e 12, que correspondem às faixas do infravermelho de ondas curtas (*SWIR*) como mostra a Figura 8.

Além da mudança das coordenadas e da radiometria, foi preciso reamostrar para 10 metros a imagem de RADAR – que tem a resolução espacial de 30 metros – para, assim, compor as bandas do Sentinel 2. Enquanto as bandas 2, 3 e 4 do visível e 8 do infravermelho próximo têm a resolução espacial de 10 metros, as bandas 11 e 12 do infravermelho de ondas curtas têm a resolução espacial de 20 metros, necessitando que fosse realizada também a reamostragem dessas bandas para 10 metros através do Aplicativo da Plataforma do Sentinel (SNAP), baixado também no sítio da Agência Espacial Europeia (ESA).



**Figura 8** - Bandas do satélite Sentinel 2  
 Fonte: Adaptado de USGS/NASA (2020)

A opção por essas bandas foi devido à variedade de objetos que foram classificados neste estudo, abrangendo maior número de refletância dos alvos. A escolha pelo satélite Sentinel 2 levou em conta: a) o tratamento dado às imagens por essas agências; b) a grande quantidade de imagens geradas em um curto período de tempo, ajudando na elaboração do mosaico com as datas das cenas próximas entre elas; c) a cobertura de extensas áreas do mundo; d) a resolução espacial mais refinada (10 metros), detectando um maior número de alvos; e) a gratuidade; e f) a acessibilidade.

#### 4.4 Seleção de amostras

Não serão coletados dados de campo para esta pesquisa. Serão explorados de forma ampla os dados apresentados nas imagens de satélite com número mínimo de amostras. A identificação das feições foi totalmente baseada na interpretação visual.


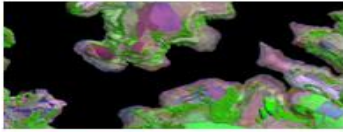



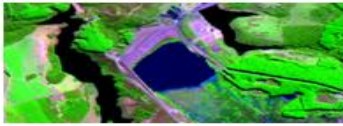







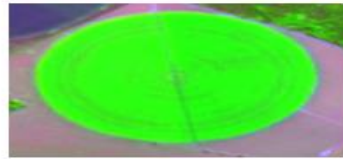



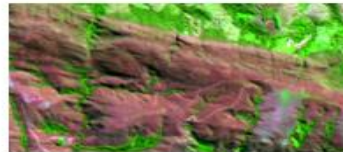


Foi criado um arquivo no formato *shape* (*ESRI Shapefile*) de polígonos para as amostras de treinamento e validação, em cuja tabela de atributos foi adicionado um novo campo com o nome das classes. O limite da área da bacia hidrográfica foi obtido

em estudos e projetos pretéritos elaborados sobre a barragem de Nova Ponte também no formato *shape*.

Os dados de alguns pontos amostrais foram levantados por observações de imagens de alta resolução espacial e compatibilidade temporal (mês de julho do ano de 2019) através da plataforma *Google Earth*. Foram gerados aleatoriamente 200 pontos amostrais que serviram para a validação da classificação. Por meio desses pontos, e também com os resultados dos dados da análise feita sobre a resposta espectral das imagens dos satélites de algumas composições de bandas produzidas no *software QGIS*, foi possível criar as classes tanto para treinamento quanto para a validação.

As composições aplicadas na imagem do Sentinel 2 foram as composições de bandas na falsa cor com as bandas do infravermelho médio, infravermelho próximo e vermelho, correspondente às bandas R-11, G-8 e B-4, e a composição também falsa cor com as bandas do infravermelho médio 1, infravermelho médio 2 e infravermelho próximo, correspondente às bandas R-12, G-11 e B-8.

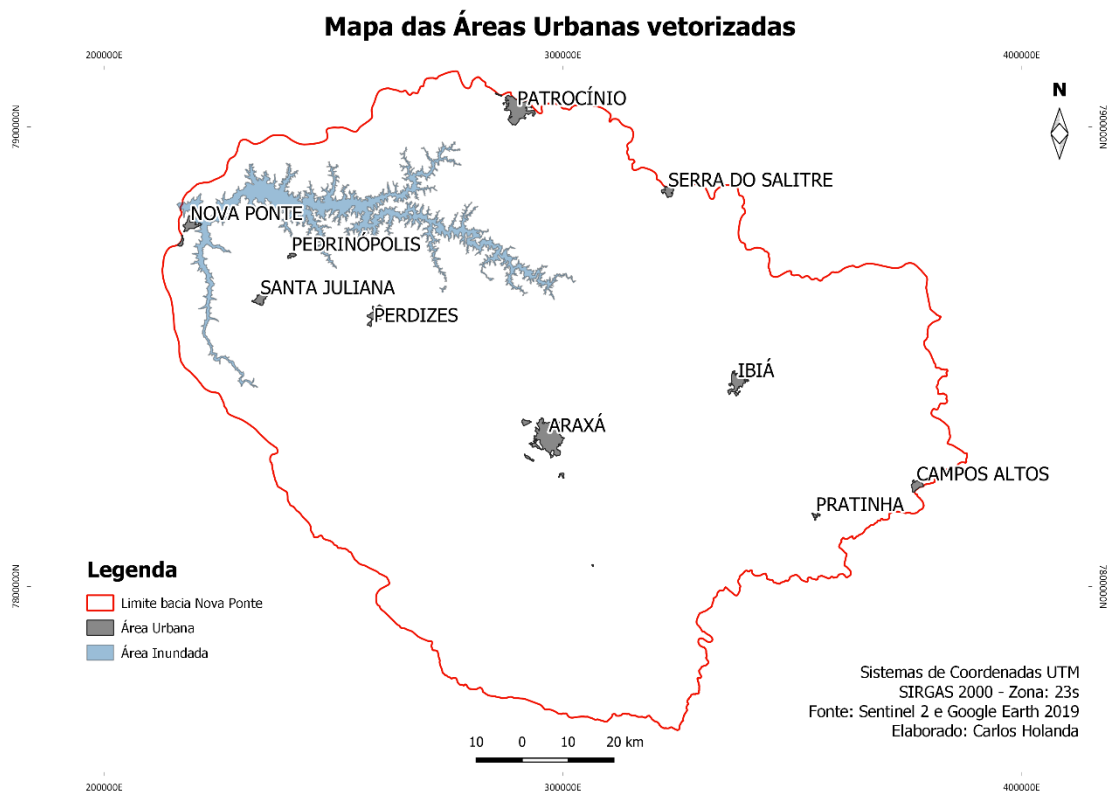
Com tais dados selecionados, foi possível identificar e determinar as amostras construindo polígonos que contemplem um conjunto de *pixels* com valores próximos entre si, permitindo caracterizá-los enquanto classe comum. Foram estabelecidas sete classes do uso e cobertura do solo para a área de abrangência do estudo: 1 - Água, cor preta; 2 - Savana Florestada, cor verde escuro com rugosidade; 3 - Água Turva, cor azul claro; 4 - Savana Rasteira Antropizada, cor roxo claro; 5 - Solo Exposto, cor branco; 6 – Agricultura Inicial, cores rosa claro e lilás claro e escuro; 7 – Agricultura Média, cor verde claro; 8 – Agricultura Avançada, cor verde escuro sem rugosidade; 9 - Savana Parque, com a cor vermelha; e 10 – Savana Arborizada, com a cor verde claro com rugosidade (Figura 9). O arquivo no formato *shape* de amostras serviu para os dois cubos de imagens formados, representando adequadamente as classes.

Classes	CHAVE DE INTERPRETAÇÃO	
	Fotos	Sentinel 2 11R 8G 4B
1. Água		
2. Savana Florestada		
3. Água Turva		
4. Savana Rasteira Antropizada		
5. Solo Exposto		
6. Agricultura Inicial		
7. Agricultura Média		
8. Agricultura Avançada		
9. Savana Parque		
10. Savana Arborizada		

**Figura 9** - Chave de Interpretação das classes

Fonte: Google Earth e Sentinel 2 (ESA).

Por possuírem muitos *pixels* misturados, os centros urbanos encontrados na área de estudo não foram selecionados como classe. Essa característica corrobora com uma maior confusão de *pixels* nessas áreas e um menor valor de índice *kappa* por meio da matriz de confusão (COLGALTON, 1999). Por meio do *software QGIS*, foram delimitados os centros urbanos criando um arquivo no formato *shape* de polígonos como mostra abaixo o mapa da Figura 10.



**Figura 10** – Mapa das Áreas Urbanas vetorizadas manualmente

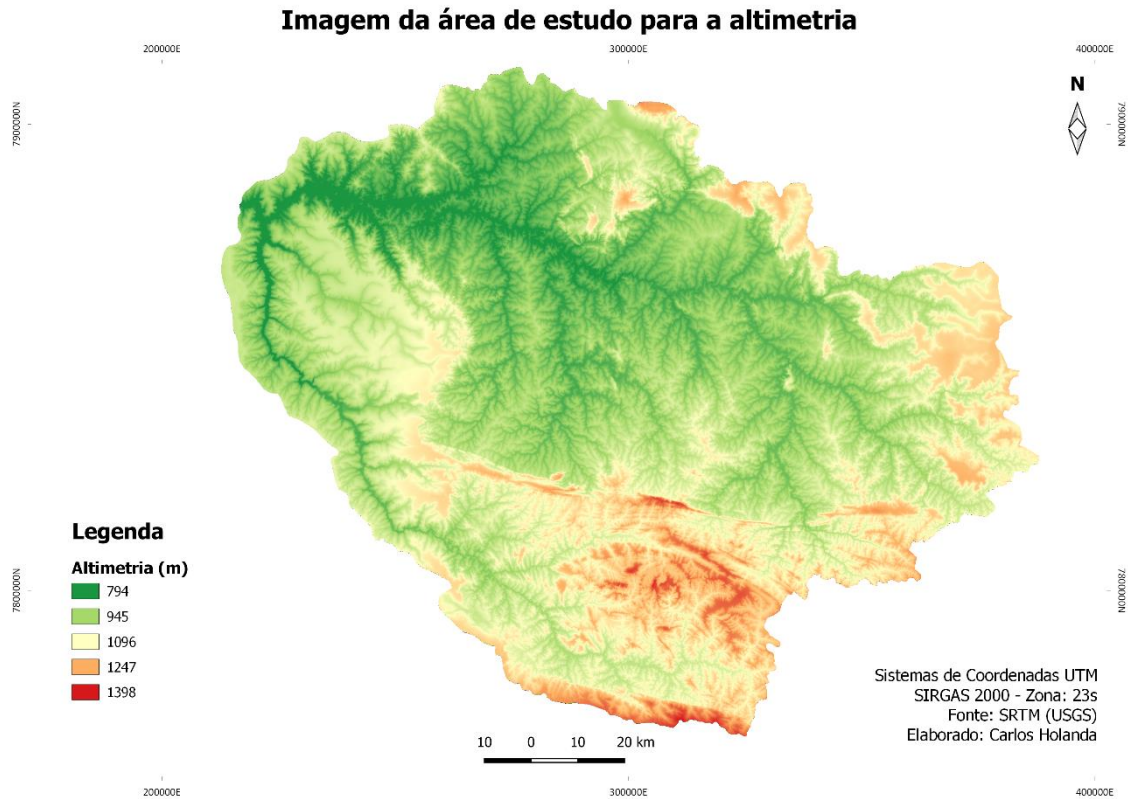
Fonte: Sentinel 2 (ESA) e Google Earth ano 2019.

## 4.5 Processamento das imagens SRTM, Declividade e Rugosidade

### 4.5.1 Imagem da área de estudo para o SRTM

A imagem mostrando a altimetria do SRTM, (Figura 11), identifica as altitudes com maiores elevações na cor vermelho escuro ou na cor laranja, e as altitudes menos elevadas na cor verde escuro e verde claro. Esta imagem serviu para compor o empilhamento de bandas junto com a imagem do satélite Sentinel 2, gerando a

classificação do uso e cobertura da terra. Para tanto, foi necessário reamostrar as imagens SRTM, declividade e rugosidade para 10 metros utilizando o *software* QGIS.



**Figura 11** – Imagem da área de estudo para a altimetria  
Fonte: SRTM (USGS).

#### 4.5.2 Imagem da área de estudo para a declividade

A imagem de declividade derivada por meio da imagem SRTM (Figura 12) para a bacia hidrográfica mostra as áreas mais planas representadas pelas cores verde escuro e verde claro – que podem ser encontradas no extremo leste e oeste da bacia nas cores verde escuro, e verde claro localizado ao centro e norte – e as áreas com maiores declividades representadas pelas cores vermelho escuro, vermelho claro e bege – nas partes mais ao sul, norte e extremo leste.



**Figura 12** - Imagem da área de estudo para a declividade  
 Fonte: SRTM (USGS).

#### 4.5.3 Imagem da área de estudo para a rugosidade

A imagem formada pelo índice de rugosidade do terreno (TRI) derivada também por meio da imagem de SRTM para a área de estudo (Figura 13) mostra os menores valores de rugosidade com as cores verde escuro e verde claro e os maiores valores de rugosidade com as cores vermelho escuro e bege. Tal qual na imagem de declividade, as evidências de alta quantidade de rugosidade foram observadas no extremo norte, no extremo sul e na parte leste da bacia hidrográfica.



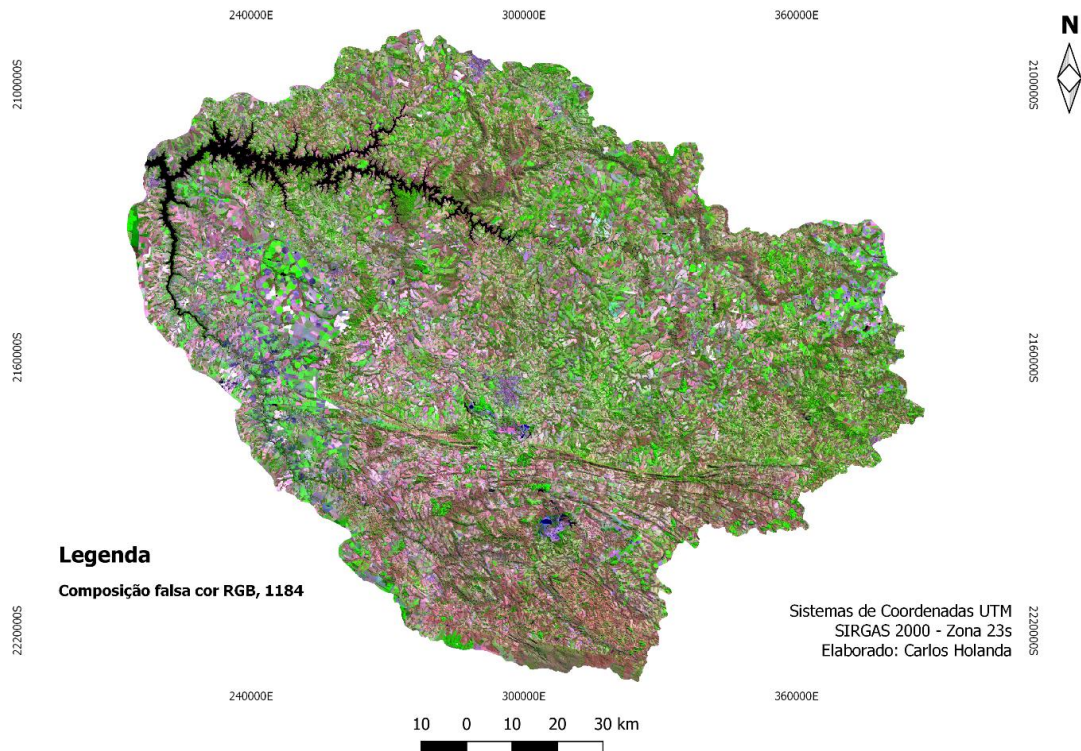
**Figura 13** - Imagem da área de estudo para a rugosidade  
Fonte: SRTM (USGS).

#### 4.5.4 Imagem empilhada para o Sentinel 2

A imagem empilhada para o satélite Sentinel 2 e suas respectivas bandas com as imagens SRTM, Declividade e Rugosidade com a resolução espacial de 10 metros segue apresentada na composição de bandas em falsa cor com R-11, G-8 e B-4 para exemplificação da mesma (Figura 14).



### Imagem mosaicada e empilhada para o Sentinel 2, SRTM, declividade e rugosidade



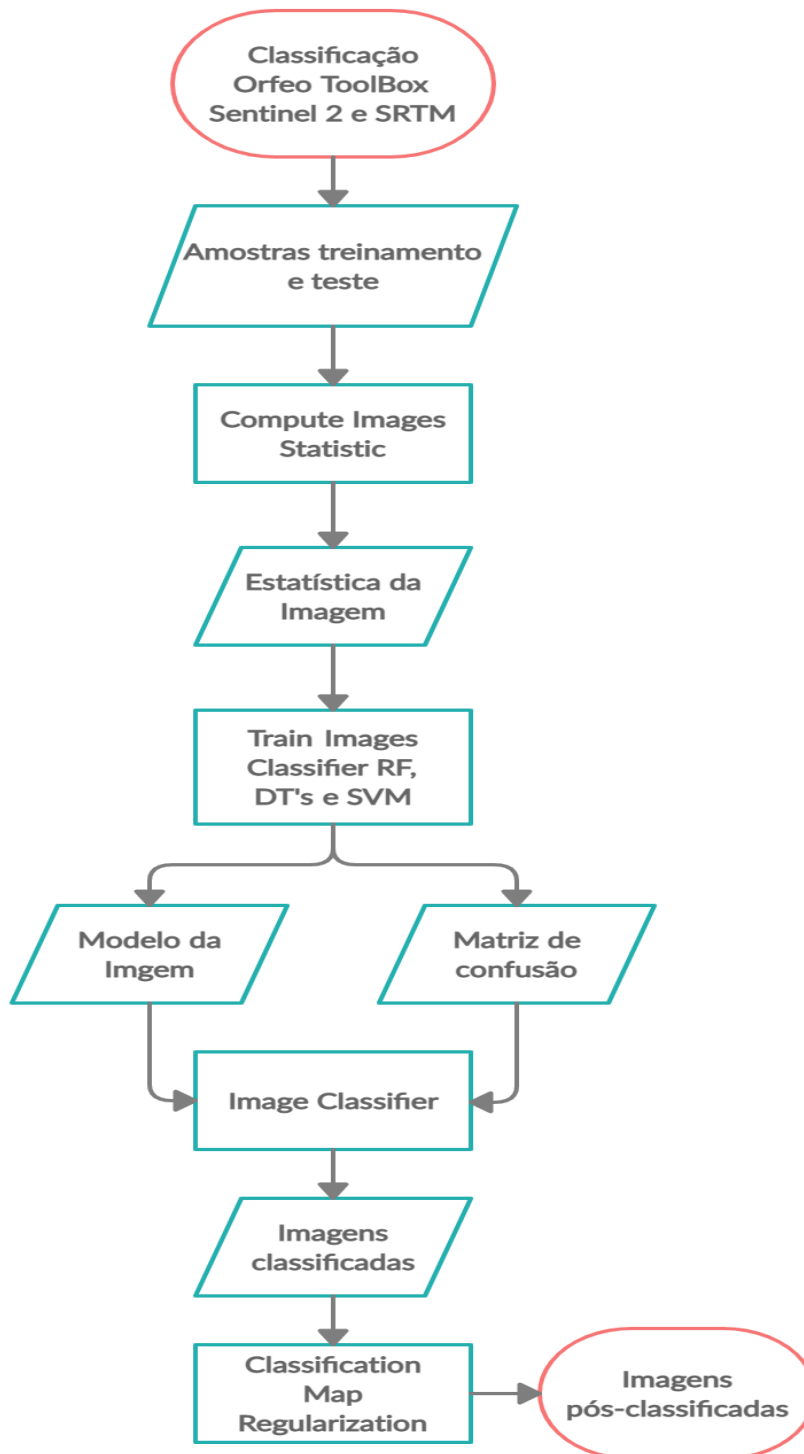
**Figura 14** - Imagem com as 6 bandas do Sentinel 2, SRTM, declividade e rugosidade  
Fonte: Sentinel 2 (ESA) e SRTM (USGS).

## 4.6 Classificação

Foram realizadas seis classificações de imagens do uso e cobertura da terra utilizando os três algoritmos de classificação mencionados para cada conjunto de imagens empilhadas, uma sem os dados altimétricos e outra com estas informações.

A classificação das imagens foi produzida no *Orfeo ToolBox* (OTB), que é um projeto de código aberto para SR que inclui um visualizador rápido de imagens (OTB, 2020). As ferramentas do projeto podem ser instaladas no *software* QGIS, tornando-se executáveis em sua interface. Alternativamente, o projeto também pode ser executado sem a utilização da interface gráfica de um SIG, demandando menos capacidade para o processamento e acelerando a conclusão das etapas.

O OTB pode ser baixado gratuitamente no sítio <<https://www.orfeo-toolbox.org/>>. No presente estudo, será utilizada a versão 7.0.0 de 64bits do projeto.



**Figura 15** - Fluxograma da classificação e pós-classificação das imagens no *Orfeo ToolBox*  
 Fonte: Autor.

Para a classificação, foi necessário gerar estatísticas como média global e desvio padrão com base na variância combinada de cada banda para os dois conjuntos de imagens por meio de ferramentas existentes no programa OTB antes de

iniciar o treinamento para cada classificador utilizado na pesquisa. Isso é preciso porque alguns algoritmos de aprendizado de máquina convergem mais rápido se a gama de recursos for  $[-1,1]$  ou  $[0,1]$ . Recursos com intervalos maiores podem ter mais peso na decisão final como no caso de algoritmos de aprendizagem de máquina que usam a distância euclidiana para compara recursos. Diante disso, é recomendável normalizar todos os recursos das imagens para a faixa  $[-1,1]$  antes de executar o treinamento (OTB, 2020). Após, será realizada a classificação das imagens. A etapa de pós-classificação tem o objetivo de uniformizar os temas, eliminando os pontos isolados classificados diferentemente de sua vizinhança, gerando uma imagem classificada com menos ruídos. Na sequência, foi preciso gerar saídas para: a) a estatística da imagem e do treinamento; b) o modelo para a imagem; e c) a matriz de confusão – usada para gerar os índices *kappa*.

Os classificadores elencados para o estudo utilizaram as amostras criadas para o treinamento e selecionadas a partir de análises feitas nas imagens do Sentinel 2 na composição falsa cor e dados das imagens do *Google Earth*. Os parâmetros utilizados nas configurações dos três classificadores do presente trabalho são mostrados na sequência. Tais parâmetros foram apresentados pela ferramenta do *software* e utilizados para os três algoritmos sem necessidade de comparação com outros parâmetros.

Os dados de amostras de treinamento e validação que foram utilizados para os três classificadores produzidos na ferramenta do *Orfeo Toolbox* para as imagens do Sentinel 2 foram (Tabela 5):

**Tabela 5 - Parâmetros para o treinamento das imagens**

Parâmetros para o treinamento das imagens	Tamanho máximo de amostras de treinamento	Tamanho máximo de validação de amostras por classe	Número de amostras vinculadas pelo mínimo	Taxa de amostra de treinamento e validação
Sentinel 2	-1	-1	1	0,5

<b>Sentinel 2 + dados altimétricos</b>	-1	-1	1	0,5
--	----	----	---	-----

Fonte: Orfeo Toolbox (OTB, 2020).

Onde:

- **Tamanho máximo da amostra de treinamento por classe valor padrão: 1000.** Tamanho máximo por classe (em pixels) da lista de amostra de treinamento (padrão = 1000) (sem limite = -1). Se for igual a -1, então o tamanho máximo da lista de amostra de treinamento disponível por classe será igual à área de superfície da menor classe multiplicada pela proporção da amostra de treinamento.
- **Tamanho máximo da amostra de validação por classe valor padrão: 1000.** Tamanho máximo por classe (em pixels) da lista de amostra de validação (padrão = 1000) (sem limite = -1). Se igual a -1, então o tamanho máximo da lista de amostra de validação disponível por classe será igual à área de superfície da menor classe multiplicada pela proporção da amostra de validação.
- **Número de amostra limitado pelo valor padrão mínimo: 1.** Limite o número de amostras para cada classe pelo número de amostras disponíveis pela classe menor. As proporções entre o treinamento e a validação são respeitadas. O padrão é verdadeiro (= 1).
- **Proporção da amostra de treinamento e validação valor padrão: 0,5.** Proporção entre as amostras de treinamento e validação (0,0 = todo o treinamento, 1,0 = toda a validação) (padrão = 0,5).

Os parâmetros que foram utilizados para os três classificadores no processo de pós-classificação (Tabela 6):

**Tabela 6:** Parâmetros para a pós-classificação das imagens

<b>Parâmetros para a Pós-classificação das imagens</b>	<b>Raio de elemento estruturador (em pixels)</b>	<b>Rótulo para a classe sem dados</b>	<b>Rótulo para a classe indecisa</b>	<b>Limite para os pixels isolados</b>
<b>Sentinel 2</b>	4	0	0	2
<b>Sentinel 2 + dados altimétricos</b>	4	0	0	2

Fonte: Orfeo Toolbox (OTB, 2020).

Onde:

- **Raio do elemento estruturante (em pixels) valor padrão: 1.** O raio do elemento estruturante em forma de bola (em pixels).
- **Rótulo para a classe sem dados valor padrão: 0.** Rótulo para a classe sem dados. Esses pixels de entrada mantêm seu rótulo sem dados na imagem de saída.
- **Rótulo para a classe Indecisa valor padrão: 0.** Rótulo para a classe Indecisos.
- **Limite para pixels isolados valor padrão: 1.** Número máximo de vizinhos com o mesmo rótulo do pixel central para considerar que se trata de um pixel isolado.

#### 4.7 Classificadores

As funcionalidades dos classificadores abordados foram detalhadas no referencial teórico, mostrando como cada um trabalha na elaboração da classificação de imagens utilizando o método de aprendizado de máquina. Nesta parte, serão apresentados os classificadores levando-se em conta os parâmetros a serem utilizados na ferramenta *Orfeo Toolbox*. Já para os parâmetros de amostras de

treinamento e validação que vão ser utilizados para a classificação para os três algoritmos a serem utilizados no OTB, tem-se:

- **Floresta Aleatória – *Random Forest (RF)***

**Tabela 7** - Parâmetros para o algoritmo *Random Forest (RF)*

Parâmetros para o Random Forest (RF)	Profundidade máxima da árvore	Número mínimo de classe em cada nó	Critérios de terminação para árvore de regressão	Agrupar valores possíveis de uma categoria variável em $k \leq \text{cat}$	Tamanho do subconjunto de recursos selecionados aleatoriamente em cada nó da árvore	Número máximo de árvores na floresta	Precisão suficiente (OOB erro)
<b>Sentinel 2</b>	5	10	0	10	0	100	0,01
<b>Sentinel 2 + dados altimétricos</b>	5	10	0	10	0	100	0,01

Fonte: Orfeo Toolbox (OTB, 2020).

Onde:

- **Profundidade máxima da árvore valor padrão: 5.** A profundidade da árvore. Um valor baixo provavelmente será insuficiente e, inversamente, um valor alto provavelmente será excessivamente adequado. O valor ideal pode ser obtido usando validação cruzada ou outros métodos adequados.
- **Número mínimo de amostras em cada nó valor padrão: 10.** Se o número de amostras em um nó for menor que este parâmetro, o nó não será dividido. Um valor razoável é uma pequena porcentagem dos dados totais, por exemplo, 1 por cento.

- **Critérios de finalização para árvore de regressão valor padrão: 0.** Se todas as diferenças absolutas entre um valor estimado em um nó e os valores das amostras de treinamento neste nó forem menores do que este parâmetro de precisão de regressão, o nó não será dividido.
  - **Agrupe os valores possíveis de uma variável categórica em clusters  $K \leq \text{cat}$  para encontrar uma divisão subótima valor padrão: 10.** Agrupe os valores possíveis de uma variável categórica em clusters  $K \leq \text{cat}$  para encontrar uma divisão subótima.
  - **Tamanho do subconjunto de recursos selecionado aleatoriamente em cada nó da árvore valor padrão: 0.** O tamanho do subconjunto de recursos, selecionado aleatoriamente em cada nó da árvore, que são usados para encontrar as melhores divisões. Se você definir como 0, o tamanho será definido como a raiz quadrada do número total de recursos.
  - **Número máximo de árvores na floresta valor padrão: 100.** O número máximo de árvores na floresta. Normalmente, quanto mais árvores você tiver, melhor será a precisão. No entanto, a melhoria na precisão geralmente diminui e atinge uma assíntota para um certo número de árvores. Também para se ter em mente, aumentar o número de árvores aumenta o tempo de previsão linearmente.
  - **Precisão suficiente (erro OOB) valor padrão: 0,01.** Precisão suficiente (erro OOB).
- **Árvores de Decisão – *Decision Tree (DT)***

**Tabela 8 - Parâmetros para o algoritmo *Decision Tree (DT)*.**

<b>Parâmetros para o <i>Decision Tree (DT)</i></b>	<b>Profundidade e máxima da árvore</b>	<b>Número mínimo de amostras em cada nó</b>	<b>Critérios de terminação para árvore de regressão</b>	<b>Agrupar valores possíveis de uma categoria variável em <math>k &lt; = \text{cat}</math></b>
<b>Sentinel 2</b>	10	10	0,01	10

<b>Sentinel 2 + dados altimétricos</b>	10	10	0,01	10
--	----	----	------	----

Fonte: Orfeo Toolbox (OTB, 2020).

Onde:

- **Profundidade máxima da árvore valor padrão: 10.** O algoritmo de treinamento tenta dividir cada nó enquanto sua profundidade é menor do que a profundidade máxima possível da árvore. A profundidade real pode ser menor se os outros critérios de terminação forem atendidos e / ou se a árvore for podada.
- **Número mínimo de amostras em cada nó valor padrão: 10.** Se o número de amostras em um nó for menor que este parâmetro, então este nó não será dividido.
- **Critérios de finalização para árvore de regressão valor padrão: 0,01.** Se todas as diferenças absolutas entre um valor estimado em um nó e os valores das amostras de treinamento neste nó forem menores do que este parâmetro de precisão de regressão, o nó não será mais dividido.
- **Agrupe os valores possíveis de uma variável categórica em clusters  $K \leq \text{cat}$  para encontrar uma divisão subótima valor padrão: 10.** Agrupe os valores possíveis de uma variável categórica em clusters  $K \leq \text{cat}$  para encontrar uma divisão subótima.

- **Máquina de Vetores de Suporte – *Support Vector Machine (SVM)***

**Tabela 9** - Parâmetros para o algoritmo *Support Vector Machine (LibSVM)*.

<b>Parâmetros para o <i>Support Vector Machines (LibSVM)</i></b>	<b>Tipo de kernel para SVM</b>	<b>Tipo de modelo de SVM</b>	<b>Parâmetro de custo C</b>	<b>Parâmetro de custo Nu</b>



<b>Sentinel 2</b>	Linear	Classificação de vetores de suporte C	1	0,5
<b>Sentinel 2 + dados altimétricos</b>	Linear	Classificação de vetores de suporte C	1	0,5

Fonte: Orfeo Toolbox (OTB, 2020).

Onde:

- **Tipo de kernel para SVM:** Linear Kernel, nenhum mapeamento é feito, esta é a opção mais rápida.
- **Tipo de modelo de SVM:** Esta formulação permite a separação imperfeita de classes. A penalidade é definida por meio do parâmetro de custo C.
- **Parâmetro de custo C Valor padrão: 1.** Os modelos SVM têm um parâmetro de custo C (1 por padrão) para controlar a compensação entre erros de treinamento e forçar margens rígidas.
- **Parâmetro de custo Nu:** Valor padrão: **0,5**. Parâmetro de custo Nu, no intervalo 0..1, quanto maior o valor, mais suave a decisão.

#### 4.8 Mapa global de uso e cobertura da terra da ESRI para o ano de 2020

O mapa global de uso e cobertura da terra da *Environmental Systems Research Institute* (ESRI) para o ano de 2020 serviu como comparação entre os dados gerados por esta pesquisa e os dados produzidos por uma empresa de alcance mundial usando conjunto de imagem igual, sem altimetria, e algoritmo apoiado em aprendizagem de máquina. A empresa americana ESRI é especializada na produção de soluções para a área de informações geográficas, disponibilizou de forma gratuita em julho de 2021 em parceria com a *Impact Observatory* e *Microsoft*, o mapa global de uso e cobertura da terra para o ano de 2020. A *Impact Observatory* baseada em Inteligência Artificial (AI) e Aprendizado de Máquina (ML) criou um modelo de classificação para o projeto utilizando como base um enorme conjunto de dados de

treinamento de bilhões de amostras disponibilizadas pelo *National Geographic Society*. Este modelo foi aplicado ao conjunto de cenas do satélite Sentinel 2, resolução espacial de 10 metros, usando seis bandas de dados de refletância de superfície: azul, verde, vermelho, infravermelho próximo e duas bandas de infravermelho de ondas curtas (KARRA *et. al.*, 2021). Este mapa está disponível em <https://www.arcgis.com>, podendo ser baixado inteiro ou em partes específicas do globo de tamanho menor.

Para este mapeamento global foram geradas dez classes de uso e cobertura da terra, mas para o recorte da área de estudo foram encontradas seis classes que são: Água, Floresta, Campo/Pasto, Vegetação Arbustiva e Solo Exposto. Como nos mapas produzidos nesta pesquisa, a classe de área urbana não foi incluída na imagem e nem na validação. Apesar deste mapa global disponibilizar os dados de validação da classificação, foi aplicado nesta imagem os mesmos índices de validação abordados nesta pesquisa.

## 5 RESULTADOS E DISCUSSÕES

Nesta pesquisa foram coletadas um total de 275 amostras para compor o treinamento dos três algoritmos como mencionado anteriormente, abrangendo uma área de 105.004.152 m<sup>2</sup> do total de 15.358.302.226 m<sup>2</sup> da área de estudo, distribuídas entre dez classes temáticas. A Tabela 10 exhibe as distribuições das amostras de treinamento para cada classe, suas estatísticas globais e as áreas totais classificadas para cada classe utilizando a imagem com os melhores índices estatísticos.

**Tabela 10** - Quadro das amostras utilizadas no treinamento dos classificadores mostrando estatísticas globais e as áreas totais classificadas para cada classe

Amostras de treinamento X Área de estudo	Números de amostras (polígonos)	Área total (m <sup>2</sup> )	Área média (m <sup>2</sup> )	Total de pixels Sentinel 2 (6 bandas) m <sup>2</sup>	Total de pixels Sentinel 2 + dados altimétricos (9 bandas) m <sup>2</sup>	Área Total classificada (ha)
--	---------------------------------	------------------------------	------------------------------	--	---	------------------------------

<b>1. Água</b>	13	15.334. 219	1.179.5 55	921.060	1.381.590	30.535,16
<b>2. Savana Florestada</b>	65	18.940. 556	295.71 9	1.142.718	1.714.077	180.224,87
<b>3. Água Turva</b>	6	2.524.4 93	420.74 9	151.590	227.385	1.467,86
<b>4. Savana Rasteira Antropizada</b>	18	2.634.6 79	150.01 3	162.180	243.270	401.718,49
<b>5. Solo Exposto</b>	33	10.128. 856	306.93 5	608.148	912.222	85.345,92
<b>6. Agricultura Inicial</b>	44	12.658. 745	287.69 9	759.798	1.139.697	97.297,67
<b>7. Agricultura Média</b>	32	22.113. 387	691.04 3	1.327.566	1.991.349	66.543,14
<b>8. Agricultura Avançada</b>	20	10.372. 926	518.64 6	623.112	934.668	27.280,12
<b>9. Savana Parque</b>	26	7.094.0 10	272.84 7	425.712	638.568	198.416,87
<b>10. Savana Arborizada</b>	18	3.202.2 81	183.52 4	198.258	297.387	447.000,14

<b>Total</b>	275	105.00 4.152	4.306.7 30	6.320.142	9.480.213	1.535.830,2 9
<b>Área de estudo</b>		15.358. 302.22 6		921.498.13 3	1.382.247.200	

Fonte: Sentinel 2 (ESA) e SRTM (USGS).

As distribuições das amostras para o treinamento seguiram uma abordagem metodológica levando-se em conta a representatividade da classe dentro da área de estudo, a proximidade geográfica e os valores de grupo de pixels limítrofes, necessitando assim de mais amostras para estas classes que obtiveram maiores confusões identificadas em processamento anteriores. Pode-se observar que a classe savana florestada teve um maior número de amostras na tentativa de diminuir a confusão desta com a classe agricultura avançada, visto que ambas possuem repostas espectrais e texturas parecidas nas composições de bandas utilizadas no processo de escolha das amostras. Isto acontece também entre as amostras das classes solo exposto e agricultura inicial. As amostras de água e água turva tiveram poucas amostras por não apresentarem muita representatividade dentro da área de estudo e pouca confusão entre as classes identificadas em processamentos anteriores.

As áreas totais das amostras de treinamento mostram suas relações com o número de polígonos de cada classe, apesar de algumas classes terem número de área maiores que outras classes, mas com menor número de amostras totais. Isso acontece também com as áreas médias das amostras totais de cada classe. O número total de pixels por área total de amostras para cada classe pode ser evidenciado nos dois conjuntos de dados tanto para a imagem Sentinel 2 com as 6 bandas espectrais quanto para o cubo de dados juntando as 6 bandas do Sentinel 2 mais as 3 bandas contendo os dados altimétricos. Vale salientar que o número de pixels encontrado para o total de cada classe foi multiplicado pelo número de bandas de cada cubo.

## 5.1 Avaliação dos Classificadores

A avaliação do desempenho dos três classificadores teve o objetivo identificar o de menor conflito entre as amostras na matriz de confusão e, por consequência, qual apresentará os melhores resultados para o índice *kappa*, para a exatidão global e para o erro de omissão e de comissão. Assim sendo, esta avaliação serviu para mostrar qual classificador tem o melhor desempenho geral, analisando estes índices estatísticos por meio da mesma imagem de satélite, e constatando se existem diferenças significativas entre classificar uma imagem apenas com as bandas disponíveis no satélite Sentinel 2, ou se existe melhora real quando empilhadas com imagens constituídas por dados altimétricos e suas derivações.

Os resultados dos algoritmos de classificação estão apresentados seguindo a ordem dos classificadores que mostraram a maior concordância entre as classes de acordo com os dados observados na matriz de confusão e, conseqüentemente, com os que tiveram os melhores índices estatísticos dentro das validações abordadas. As Tabelas 11 e 12 mostram os resultados obtidos neste estudo para os algoritmos abordados tanto para os índices estatísticos quanto para o tempo gasto de processamento.

**Tabela 11** - Quadro comparativo dos resultados *kappa* obtidos nos classificadores

<b>Resultados <i>kappa</i> para as imagens e classificadores</b>	<b>Algoritmo SVM</b>	<b>Algoritmo DT</b>	<b>Algoritmo RF</b>
<b>Sentinel 2</b>	0,89	0,87	0,74
<b>Sentinel 2 + dados altimétricos</b>	0,91	0,89	0,76

Fonte: Orfeo Toolbox (OTB, 2020).

**Tabela 12** - Quadro comparativo do tempo gasto do processamento dos dados de treinamento, classificação e pós-classificação para os algoritmos

<b>Resultados do tempo gasto para processar os dados</b>	<b>Algoritmo SVM</b>	<b>Algoritmo DT</b>	<b>Algoritmo RF</b>
<b>Sentinel 2</b>	48 horas seguidas	40 minutos seguidos	50 minutos seguidos
<b>Sentinel 2 + dados altimétricos</b>	48 horas seguidas	40 minutos seguidos	50 minutos seguidos

Fonte: Orfeo Toolbox (OTB, 2020).

Os algoritmos abordados foram eficientes na identificação das classes encontradas quando aplicadas técnicas de processamento digital de imagens (PDI) e sensoriamento remoto (SR), aliadas com a aprendizagem de máquina. O classificador que obteve os melhores índices estatísticos foi o SVM tanto para as imagens com os dados altimétricos, com *kappa* de 0,91, quanto para as imagens sem estas informações complementares, com *kappa* de 0,89. Apesar do DT ter apresentado resultados inferiores ao SVM, o classificador de DT obteve tempo de processamento para treinamento, classificação e pós-classificação de 40 minutos, muito menor que o do SVM.

Diante desse dado, observa-se que, caso a área de pesquisa seja muito extensa ou tenha várias pequenas áreas aliadas com um computador com processador não adequado, o classificador SVM não seria a melhor opção. O algoritmo RF obteve praticamente o mesmo tempo de processamento que o DT, mas atingiu os menores índices estatísticos entre os três classificadores. A utilização do cubo de dados com a imagem do SRTM e suas derivações como declividade e rugosidade na classificação do uso e cobertura da terra para os três algoritmos mostraram resultados superiores em comparação com as imagens sem estes dados altimétricos.

Alguns resultados comparando algoritmos de classificação baseados em aprendizado de máquina podem ser observados em Cavalcanti (2016). Nesta pesquisa o autor trabalhou com imagens de pequenas áreas da região da Amazônia Legal geradas por meio de aviões que possuem estes sensores ópticos ativo, (PDI) e

aprendizado de máquina para detectar manchas de ação antrópica dentro da floresta. Apesar de empregar metodologia distinta como segmentação de imagem e extração de atributos das imagens como morfologia, cor, textura, entre outros, ambas empregando técnicas matemáticas tendo como base os valores dos pixels. O estudo comparou o desempenho de algoritmos de classificação fundamentados em aprendizado de máquina em dois conjuntos de dados como também está sendo aplicado para esta pesquisa. O primeiro conjunto de dados foi classificado utilizando todo o montante de atributos extraídos da imagem. Para a segunda classificação foi utilizado a técnica *Correlation-based Feature Selection (CFS)*, introduzida por Hall (1998) para selecionar entre o montante os atributos que possuem alta correlação com as classes das amostras.

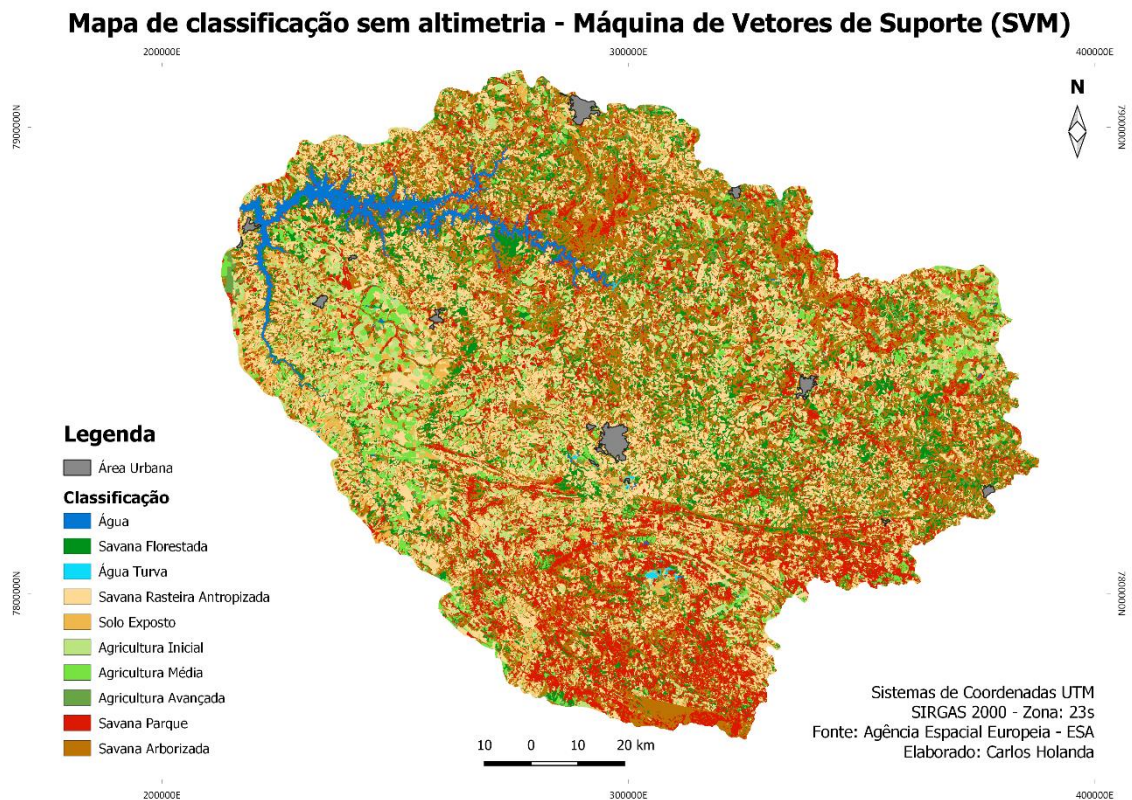
Os resultados encontrados por Cavalcanti (2016) apresentam a precisão e a acurácia da classificação de imagens para diversos algoritmos baseados em aprendizagem de máquina para estes dois conjuntos de dados mencionados. Apesar da grande quantidade de atributos nos dois conjuntos de dados, as imagens utilizadas possuem apenas três bandas, vermelho, verde e azul (RGB). Este número limitado de bandas foi empregado possivelmente por não existir imagens da área de estudo disponíveis de alta resolução espacial e com número de bandas maior.

Ao contrário do foi encontrado nesta pesquisa, o estudo citado mostrou que o algoritmo de RF obteve o melhor resultado no desempenho, provavelmente por se tratar de imagem segmentada e imagens com resolução espacial mais refinada. Assim sendo, a pesquisa atual vem contribuir e acrescentar conforme os dados citados no estudo acima, novas informações sobre a avaliação de algoritmos de aprendizado de máquina na classificação de imagens, utilizando um maior número de bandas ou respostas espectrais dos objetos. Atribuindo também a estas bandas informações adicionais de altimetria, declividade e rugosidade.

### **5.1.1 Imagem classificada sem altimetria para o algoritmo de *Support Vector Machine (SVM)***

O mapa de classificação do uso e cobertura da terra sem altimetria para o algoritmo Máquina de Vetores de Suporte (SVM) pode ser visualizado na Figura 16 e

mostra a distribuição de suas respectivas classes dentro da área de estudo. Na imagem também é possível notar as áreas urbanas desenhadas em polígonos na cor cinza, conforme a legenda da figura. O algoritmo de SVM mostrou ser, dentre os classificadores apoiados em aprendizado de máquina testados nesta pesquisa, o que obteve o melhor resultado para a área de estudo para a imagem sem os dados altimétricos e derivações.



**Figura 16** - Mapa de classificação sem altimetria – Máquina de Vetores de Suporte (SVM)  
 Fonte: Agência Espacial Europeia - ESA.

Por meio dos dados da pesquisa, pode-se observar que o classificador conseguiu distribuir as classes dentro da área da bacia hidrográfica de acordo com as amostras de treinamento e validação, bem como o que foi visualizado nas imagens do Google Earth para o ano de 2019. O algoritmo abordou toda a região do extremo sul, norte e nordeste como mostra a Figura 16, na cor vermelha para a savana parque, e na cor marrom para a savana arborizada, a qual configura as áreas altimétricas mais elevadas, alcançando uma parte da Serra da Canastra.



A classe água mostrou toda a área alagada pela barragem de Nova Ponte, enquanto que as pequenas áreas foram demonstradas por meio da classe água turva, como bacias de rejeitos de algumas mineradoras existentes no extremo sul, bem como pequenos lagos construídos próximos dos pivôs centrais, possivelmente para o seu abastecimento.

Uma grande quantidade de regiões com savana rasteira antropizada, na cor bege claro, e com solo exposto, na cor bege escuro, foi notada dentro do limite estudado com áreas representativas espalhadas em quase todas as áreas da bacia, mas com ênfase no centro e oeste dela. Pode-se observado, também, que o classificador conseguiu abordar as áreas de agricultura nos seus diversos estágios evolutivos, evidenciadas, em sua maior parte, no extremo leste e oeste da imagem, mostrando predomínio da agricultura.

Em contrapartida às classes anteriores aqui citadas, por meio das quais é perceptível a alteração pelo homem, a savana florestada, na cor verde escuro, não apresenta muitas áreas na bacia, sendo evidenciada, em sua maior parte, perto da área alagada, ao norte, e alguns fragmentos no extremo leste.

Os resultados da matriz de confusão com valores representados em números de pixels para o classificador Máquina de Vetores de Suporte com a imagem sem altimetria são apresentados na Tabela 13 e demonstram as confusões entre as classes definidas para este estudo.

**Tabela 13** - Matriz de confusão (pixels), erro de omissão e comissão, exatidão global e índice *kappa* para a Máquina de Vetores de Suporte (SVM) sem altimetria

<b>Linhas e Colunas</b>	<b>Água</b>	<b>Savana Florestada</b>	<b>Água Turva</b>	<b>Savana Rasteira Antropizada</b>	<b>Solo Exposto</b>	<b>Agricultura Inicial</b>	<b>Agricultura Média</b>	<b>Agricultura Avançada</b>	<b>Savana Parque</b>	<b>Savana Arborizada</b>	<b>Total</b>
<b>Água</b>	54480	0	0	0	0	0	0	0	0	0	54480
<b>Savana Florestada</b>	0	11184	0	1	0	0	188	407	1	683	12464
<b>Água Turva</b>	0	0	7728	0	0	0	0	0	0	0	7728
<b>Savana Rasteira Antropizada</b>	0	21	0	32510	832	76	119	3	577	3392	37530
<b>Solo Exposto</b>	0	0	0	173	12151	0	0	0	0	0	12324
<b>Agricultura Inicial</b>	0	0	0	4	1955	20213	0	0	2	0	22174

<b>Agricultura Média</b>	0	329	0	0	0	0	14911	1	0	12	15253
<b>Agricultura Avançada</b>	0	5180	0	0	0	0	58	11440	0	462	17140
<b>Savana Parque</b>	0	3	0	1676	0	34	429	0	40374	2874	45390
<b>Savana Arborizada</b>	0	898	0	366	0	0	1	16	2848	17281	21410
<b>Total</b>	54480	17615	7728	34730	14938	20323	15706	11867	43802	24704	245893
<b>Erro de Comissão</b>	0%	36,5%	0%	6,3%	18,3%	0,5%	5%	3,5%	7,8%	30%	
<b>Erro de Omissão</b>	0%	10,2%	0%	13,3%	1,4%	8,8%	2,2%	33,2%	11%	19,2%	
<b>Exatidão Global</b>	0,90										
<b>Índice Kappa</b>	0,89										

Fonte: Orfeo Toolbox (OTB, 2020).

As classes que tiveram as maiores confusões para os dados citados foram: a agricultura inicial, com um total de 22.174 pixels, em que 1.955 foram confundidos com o solo exposto; a agricultura avançada, com um total de 17.140 pixels, dentre os quais 5.180 foram confundidos com savana florestada; a savana rasteira antropizada com total de 37.530 pixels, dos quais 3.392 foram confundidos com a savana arborizada; a savana parque, com 45.390, dos quais 1.676 foram confundidos com a savana rasteira antropizada e 2.874 confundidos com a savana arborizada; além da savana arborizada, que dos 21.410 pixels apresentados, 2.848 foram confundidos com a savana parque. As classes que não tiveram nenhuma ou pouca confusão na matriz foram: água, água turva, solo exposto e agricultura média. A respeito da ocorrência dessas confusões entre as classes evidenciadas na matriz, acredita-se que elas se dão possivelmente por causa dos limites estabelecidos entre uma determinada classe e sua vizinhança e por ambas possuírem valores de refletâncias e números digitais constituídos para cada pixel da imagem próximos.

Após a observação desses dados, o índice *kappa* foi gerado a partir do resultado da matriz de confusão. Apesar das confusões apresentadas entre algumas classes, este classificador mostrou ter bons resultados na hora de classificar uma imagem, obtendo um índice *kappa* de 0,89 e, assim, mostrando que está dentro de um nível de concordância excelente, conforme Tabela 2.

Por sua vez, o índice de exatidão global mostrou um resultado satisfatório de 0,90, acima do resultado apresentado pelo índice *kappa* para esta classificação de imagem e com este classificador. Todavia, conforme Brites (1996), o índice de exatidão global não abarca os erros de omissão e comissão contidos na matriz de confusão, superestimando a avaliação da exatidão e resultando sempre em valores mais altos que os do índice *kappa*, o qual abrange todos os dados da matriz de confusão.

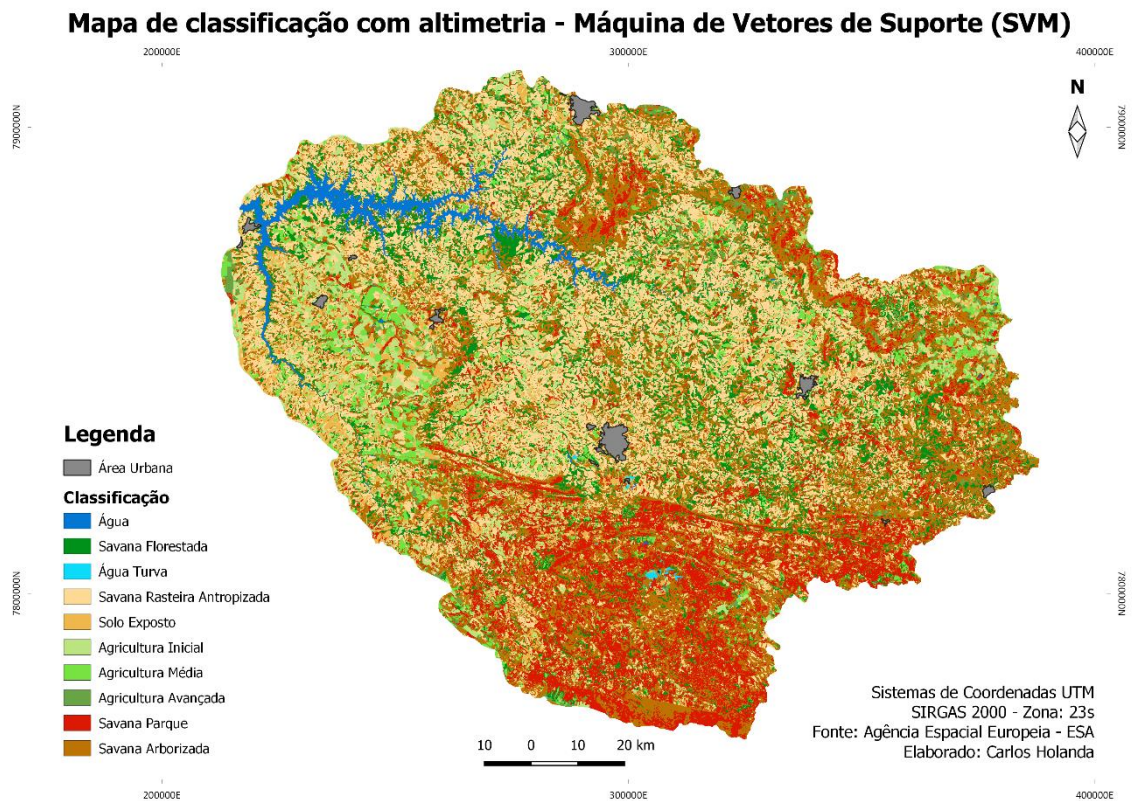
Os resultados para o erro de omissão, que são os números de pixels que não foram incluídos em determinada categoria, mostraram os maiores erros para as classes como agricultura avançada, com resultado de 33,2% do total; e savana parque, com 19,2%. Para os resultados do erro de comissão, os pixels da imagem que também não foram incluídos em nenhuma classe na diagonal da matriz de confusão

mostraram os maiores dados para as classes savana florestada, com resultado de 36,5%; solo exposto, com 18,3%; e savana arborizada, com 30% do total.

Quanto ao tempo gasto para a realização do processamento dos dados da classificação e pós-classificação desta imagem para este classificador, conforme evidenciado na Figura 16, que mostra o fluxograma da classificação, cumpre lembrar que devem ser somadas: a) Estatística da imagem; b) Treinamento da imagem com o classificador, em que também são gerados o modelo e a matriz de confusão; c) Classificação de imagem; e d) Pós-classificação da imagem. Todo esse processo teve a duração de, aproximadamente, 48 horas seguidas.

### **5.1.2 Imagem classificada com altimetria para o algoritmo de *Support Vector Machine* (SVM)**

O mapa de classificação do uso e cobertura da terra da imagem, agora com os dados de altimetria para o algoritmo Máquina de Vetores de Suporte (SVM), pode ser visualizado na Figura 17. O algoritmo de SVM mostrou ser, dentre os classificadores apoiados em aprendizado de máquina testados nesta pesquisa, o que obteve o melhor resultado para a área de estudo com relação à imagem com os dados altimétricos e derivações entre os três algoritmos abordados, apresentando também melhores resultados na classificação comparada com o mesmo classificador, mas com a imagem sem altimetria.



**Figura 17** - Mapa de classificação com altimetria – Máquina de Vetores de Suporte (SVM)  
Fonte: Agência Espacial Europeia - ESA.

Na classificação com altimetria por meio do algoritmo Máquina de Vetores de Suporte (SVM), o classificador mostrou melhoras ao distribuir as classes dentro da área de estudo utilizando o conjunto de dados altimétricos. Dessa maneira, foi possível visualizar de forma mais nítida as partes mais planas do relevo e as mais elevadas.

Embora haja maior definição das classes savana parque e savana arborizada para esta imagem em relação à anterior, estas classes ainda podem ser visualizadas, em sua maior parte, no extremo sul como mostra a Figura 17, no extremo norte e nordeste na cor vermelha para a savana parque e na cor laranja para a savana arborizada. A classe água não modificou o seu formato em comparação com a imagem anterior, assim como as pequenas áreas classificadas como água turva.

A grande quantidade de regiões com savana rasteira antropizada, na cor bege claro, e solo exposto, na cor bege escuro, continuaram representativas, encontrando-se espalhadas em quase todas as áreas da bacia. Todavia, com o auxílio dos dados altimétricos, essas regiões ficaram mais definidas principalmente no centro da imagem, além de ser possível melhor identificar áreas destinadas à agricultura nas

partes mais planas da bacia, evidenciadas, em sua maior parte, no extremo leste e oeste da imagem, nas cores entre verde claro, verde limão e verde musgo. A classe savana florestada na cor verde escuro, continua caracterizada majoritariamente perto da área alagada ao norte e alguns fragmentos no extremo leste e centro-oeste.

Os resultados da matriz de confusão em valores de pixels para o classificador Máquina de Vetores de Suporte, com a imagem utilizando os dados de altimetria apresentados na Tabela 14, demonstram as confusões entre as classes definidas para este estudo a seguir:

**Tabela 14** - Matriz de confusão (pixels), erro de omissão e comissão, exatidão global e índice kappa para a Máquina de Vetores de Suporte (SVM) com altimetria

<b>Linhas e Colunas</b>	<b>Água</b>	<b>Savana Florestada</b>	<b>Água Turva</b>	<b>Savana Rasteira Antropizada</b>	<b>Solo Exposto</b>	<b>Agricultura Inicial</b>	<b>Agricultura Média</b>	<b>Agricultura Avançada</b>	<b>Savana Parque</b>	<b>Savana Arborizada</b>	<b>Total</b>
<b>Água</b>	54480	0	0	0	0	0	0	0	0	0	54480
<b>Savana Florestada</b>	0	11252	0	0	0	0	110	326	0	776	12464
<b>Água Turva</b>	0	0	7728	0	0	0	0	0	0	0	7728
<b>Savana Rasteira Antropizada</b>	0	21	0	33029	714	38	15	0	380	3333	37530
<b>Solo Exposto</b>	0	0	0	85	12239	0	0	0	0	0	12324
<b>Agricultura Inicial</b>	0	0	0	1	2181	19964	0	0	28	0	22174



<b>Agricultura Média</b>	0	318	0	0	0	0	14933	2	0	0	15253
<b>Agricultura Avançada</b>	0	2182	0	0	0	0	0	14733	0	225	17140
<b>Savana Parque</b>	0	0	0	1051	0	199	0	0	40796	3344	45390
<b>Savana Arborizada</b>	0	774	0	82	0	0	0	1	2762	17791	21410
<b>Total</b>	54480	14547	7728	34248	15134	20201	15058	15062	43966	25469	245893
<b>Erro de Comissão</b>	0%	22,6%	0%	3,5%	19,1%	1,1%	0,8%	2,1%	7,2%	30,1%	
<b>Erro de Omissão</b>	0%	9,7%	0%	11,9%	0,6%	9,9%	3,4%	14%	10,1%	16,9%	
<b>Exatidão Global</b>	0,92										
<b>Índice Kappa</b>	0,91										

Fonte: Orfeo Toolbox (OTB, 2020).

Dessa forma, nota-se que a matriz de confusão da classificação com os dados altimétricos mostrou resultados melhores em relação aos dados sem altimetria, em que é possível observar menores confusões entre as classes na matriz. As classes que tiveram as maiores confusões para os dados citados foram a savana rasteira antropizada, com um total de 37.530 pixels, dentre os quais 3.333 foram confundidos com savana arborizada; a agricultura inicial, com 22.174 pixels, dos quais 2.181 foram considerados como solo exposto; a agricultura avançada, com um total de 17.140 pixels, em que 2.182 foram confundidos com savana florestada; a savana parque, com 45.390 pixels, dentre os quais 1.051 foram confundidos com savana rasteira antropizada e 3.344 foram confundidos com savana arborizada; e as maiores confusões para a classe savana arborizada, que, do total de 21.410 pixels, 2.762 foram confundidos com savana parque. As classes que não tiveram nenhuma ou pouca confusão na matriz continuaram sendo as classes: água, água turva, solo exposto e agricultura média. Estas confusões entre as classes evidenciadas na matriz possivelmente foram por causa dos limites estabelecidos entre as mesmas e sua vizinhança, e também por possuírem valores de refletâncias e números digitais constituídos para cada pixel da imagem próximos.

Nesse sentido, a partir do resultado da matriz de confusão, foi gerado o índice *kappa*. Apesar das confusões apresentadas entre algumas classes, este classificador mostrou ter o melhor resultado para a imagem composta com os dados altimétricos entre os classificadores abordados no estudo para a classificação do uso e cobertura da terra, obtendo-se um índice *kappa* de 0,91, que evidencia estar dentro do nível de concordância considerado excelente, conforme a Tabela 2.

O índice de exatidão global também mostrou um resultado satisfatório de 0,92, acima do resultado apresentado pelo índice *kappa* para esta classificação de imagem, que utilizou o empilhamento dos dados com informações altimétricas e derivações, superando também a exatidão global do mesmo algoritmo, mas contando apenas com as informações do satélite Sentinel 2.

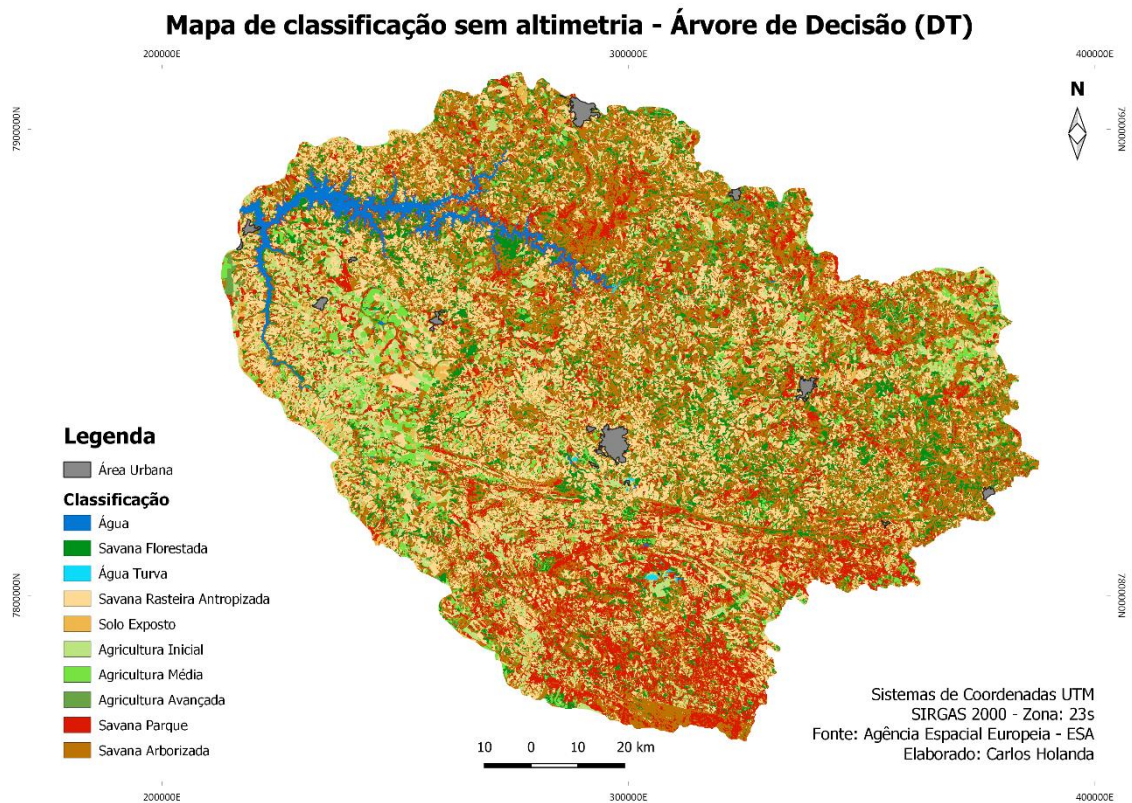
Os resultados para o erro de omissão mostraram um maior erro encontrado, que foi para as classes agricultura avançada, com resultado de 14% e savana arborizada, com 16,9%. Já para os resultados do erro de comissão, foram obtidos os

maiores dados para as classes savana arborizada, com o resultado de 22,6%; solo exposto, com 19,1%; e savana arborizada, com 30,1% do total.

Quanto ao tempo gasto para a realização do processamento dos dados da classificação e pós-classificação desta imagem para este classificador, somam-se, também: a) Estatística da imagem; b) Treinamento da imagem com o classificador, em que também são gerados o modelo e a matriz de confusão; c) Classificação de imagem; e d) Pós-classificação da imagem, que foi de aproximadamente 48 horas seguidas, mesmo utilizando-se a imagem com os dados complementares de altimetria.

### **5.1.3 Imagem classificada sem altimetria para o algoritmo de Árvore de Decisão (DT)**

O mapa de classificação do uso e cobertura da terra sem altimetria para o algoritmo Árvore de Decisão (DT) pode ser visualizado na Figura 18. Nesta análise, o algoritmo de DT mostrou ser, dentre os classificadores apoiados em aprendizado de máquina testados nesta pesquisa, o que obteve o resultado inferior em comparação ao SVM para os dados sem altimetria e derivações para a área de estudo, superando apenas o desempenho do algoritmo de Floresta Aleatória (RF) para o mesmo conjunto de dados.



**Figura 18 - Mapa de classificação sem altimetria – Árvore de Decisão (DT)**  
Fonte: Agência Espacial Europeia - ESA.

No âmbito da classificação sem altimetria com a Árvore de Decisão (DT), o algoritmo também conseguiu abordar toda a região sul da bacia hidrográfica mencionada anteriormente para o classificador SVM, configurada como savana parque e savana arborizada. Estas classes ainda podem ser visualizadas, em sua maior parte, no extremo sul da imagem, no extremo norte e nordeste pela cor vermelha para a savana parque, e cor marrom para a savana arborizada. Esta pode ser evidenciada, ainda, no extremo leste e centro, mostrando uma maior quantidade em área visualmente da savana arborizada e savana parque em relação à classificação do SVM sem altimetria.

Para a classe água, o classificador continuou mostrando toda a área alagada pela barragem de Nova Ponte, como também as pequenas áreas com a classe água turva. Observando a imagem classificada e os dados contidos na matriz de confusão, nota-se que estas duas classes não sofreram muita alteração entre os classificadores abordados.

Por fim, o classificador mostrou, como os anteriores, a grande quantidade de regiões com savana rasteira antropizada, na cor bege claro, e solo exposto, na cor bege escuro, para a área de estudo. Nota-se que, dentro do limite estudado, estas classes estão espalhadas em quase todas as áreas da bacia, mas com ênfase no centro e oeste da mesma. Também é possível notar que o classificador conseguiu abordar as áreas de agricultura evidenciadas em sua maior parte no extremo leste e oeste da imagem, reforçando que esta é uma área que tem sua base econômica no agronegócio.

Visualmente, também se nota que esta classe teve uma menor quantidade em área comparada com o classificador SVM sem altimetria. A savana florestada, na cor verde escuro, não apresentou muitas mudanças dentro da área na bacia comparada com as classificações anteriores, sendo evidenciada ao norte e alguns fragmentos no extremo leste.

Os resultados da matriz de confusão (pixels) para o classificador Árvore de Decisão com a imagem sem altimetria são apresentados na Tabela 15 e demonstram as confusões entre as classes definidas para este estudo.

**Tabela 15** - Matriz de confusão (pixels), erro de omissão e comissão, exatidão global e índice *kappa* para a Árvore de Decisão (DT) sem altimetria

<b>Linhas e Colunas</b>	<b>Água</b>	<b>Savana Florestada</b>	<b>Água Turva</b>	<b>Savana Rasteira Antropizada</b>	<b>Solo Exposto</b>	<b>Agricultura Inicial</b>	<b>Agricultura Média</b>	<b>Agricultura Avançada</b>	<b>Savana Parque</b>	<b>Savana Arborizada</b>	<b>Total</b>
<b>Água</b>	54480	0	0	0	0	0	0	0	0	0	54480
<b>Savana Florestada</b>	0	10581	0	1	0	0	140	886	1	855	12464
<b>Água Turva</b>	0	0	7728	0	0	0	0	0	0	0	7728
<b>Savana Rasteira Antropizada</b>	0	38	0	31113	1082	184	98	13	611	4391	37530
<b>Solo Exposto</b>	0	0	0	144	12180	0	0	0	0	0	12324
<b>Agricultura Inicial</b>	0	0	0	76	901	21194	0	0	0	3	22174

<b>Agricultura Média</b>	0	363	0	19	0	0	14688	1	0	182	15253
<b>Agricultura Avançada</b>	0	5405	0	0	0	0	72	10487	0	1176	17140
<b>Savana Parque</b>	0	3	0	1552	236	724	0	0	38504	4371	45390
<b>Savana Arborizada</b>	0	722	0	363	7	3	5	108	2920	17282	21410
<b>Total</b>	54480	17112	7728	33268	14406	22105	15003	11495	42036	28260	245893
<b>Erro de Comissão</b>	0%	38,1%	0%	6,4%	15,4%	4,1%	2%	8,7%	8,4%	38,8%	
<b>Erro de Omissão</b>	0%	15,1%	0%	17%	1,16%	4,4%	3,7%	38,8%	15,1%	19,2%	
<b>Exatidão Global</b>	0,89										
<b>Índice Kappa</b>	0,87										

Fonte: Orfeo Toolbox (OTB, 2020).

As classes que tiveram as maiores confusões para os dados citados foram a agricultura avançada, com um total de 17.140 pixels, dos quais 5.405 foram confundidos com savana florestada e 1.176 com savana arborizada; a savana rasteira antropizada, com 37.530 pixels, em que 1.082 foram confundidos com solo exposto e 4.391, com a savana arborizada; a savana parque, que de 45.390 pixels, dos quais 1.552 foram para a savana rasteira antropizada e 4.351 para a savana arborizada; e, por fim, dos 21.410 pixels da savana arborizada, 2.920 foram para a savana parque.

As classes que continuaram com nenhuma ou pouca confusão na matriz foram: água, água turva, solo exposto e agricultura média. Estas confusões entre as classes evidenciadas na matriz possivelmente se deram por causa dos limites estabelecidos entre a mesma e sua vizinhança, e por possuírem valores de refletâncias e números digitais constituídos para cada pixel da imagem próximos.

Também nesta etapa, a partir do resultado da matriz de confusão, foi gerado o índice *kappa*. Apesar das confusões apresentadas entre algumas classes, este instrumento mostrou ter bons resultados na hora de classificar uma imagem, obtendo um índice *kappa* de 0,87, que revela estar dentro do nível de concordância excelente, conforme Tabela 2.

De forma a agregar, o índice de exatidão global mostrou resultado satisfatório de 0,89, acima do resultado apresentado pelo índice *kappa* para esta classificação de imagem, que utilizou os dados sem as informações altimétricas e derivações.

Os resultados para o erro de omissão mostraram maiores erros para as classes como agricultura avançada, com resultado de 38,8%; savana florestada, com 15,1%; savana parque, com 15,1%; e savana arborizada, com 19,2% do total. Esses achados corroboram o que a matriz de confusão apresentou na Tabela 12.

Já os resultados do erro de comissão mostraram maiores dados para as classes savana florestada, com o resultado de 38,1%; solo exposto, com 15,4%; e savana arborizada, com 38,8%.

Quanto ao tempo gasto para a realização do processamento dos dados da classificação e pós-classificação desta imagem para este classificador, somam-se, também: a) Estatística da imagem; b) Treinamento da imagem com o classificador, em que também são gerados o modelo e a matriz de confusão; c) Classificação de

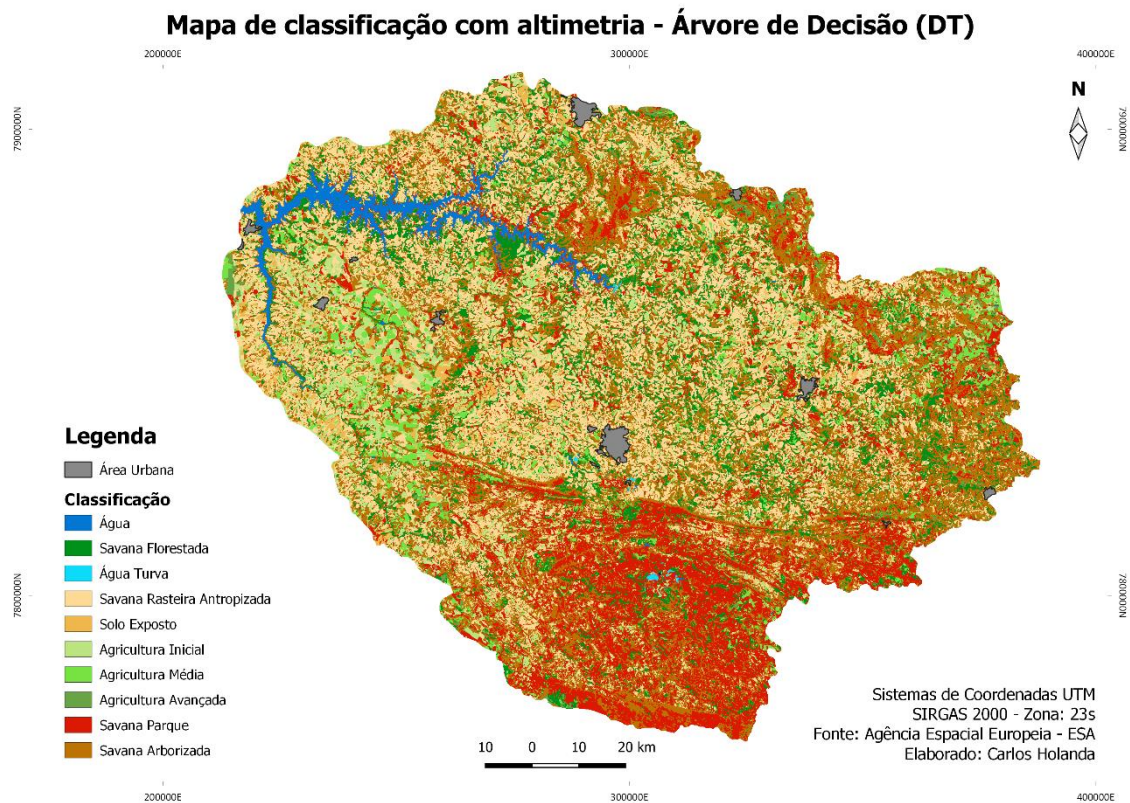


imagem; e d) Pós-classificação da imagem, que foi de aproximadamente 40 minutos seguidos.

#### **5.1.4 Imagem classificada com altimetria para o algoritmo de Árvore de Decisão (DT)**

O mapa de classificação do uso e cobertura da terra com altimetria para o algoritmo Árvore de Decisão (DT) pode ser visualizado na Figura 19. O algoritmo de DT para a imagem empilhada com os dados altimétricos e derivações mostrou como o algoritmo anterior (SVM), que obteve melhoras nos resultados observando as validações utilizadas para esta pesquisa, quando empregadas informações complementares ao cubo de dados na composição da imagem para a classificação, se distingue dos demais.

Todavia, esse algoritmo não conseguiu os melhores desempenhos para os três classificadores apoiados em aprendizagem de máquina para a área de estudo nesta pesquisa, o que faz com ele perca para o SVM, embora tenha conquistado resultados melhores que o algoritmo de Floresta Aleatória (RF), apresentado mais adiante.



**Figura 19 - Mapa de classificação com altimetria – Árvore de Decisão (DT)**  
Fonte: Agência Espacial Europeia - ESA.

Assim, na classificação com altimetria em DT, o classificador mostrou melhoras ao distribuir as classes dentro da área de estudo utilizando o conjunto de dados altimétricos, podendo assim, também como no algoritmo SVM – com altimetria –, visualizar de forma mais nítida as partes mais planas do relevo, bem como as mais elevadas. Pode-se notar que essa ferramenta também conseguiu abordar toda a região sul da bacia hidrográfica, como na imagem anterior sem altimetria, mas agora com uma definição mais precisa das maiores cotas altimétricas classificadas como savana parque e savana arborizada.

As cores vermelho (savana parque) e marrom (savana arborizada) auxiliam a visualizar as áreas destacadas, em sua maior parte, no extremo sul da imagem e no extremo norte e nordeste. Estas classes, em relação à imagem anterior sem altimetria e à classificação com o algoritmo SVM com os dados altimétricos, mostraram visualmente o aumento em quantidade de área, principalmente no extremo sul da imagem para a savana parque, e no leste da imagem para a savana arborizada.

A classe água, todavia, não modificou o seu formato em comparação com as imagens anteriores, como também as pequenas áreas com a classe água turva.

A grande quantidade de regiões com savana rasteira antropizada (em bege claro), e solo exposto (bege escuro), continuaram representativas e espalhadas em quase todas as áreas da bacia. Entretanto, com o auxílio dos dados altimétricos, elas ficaram mais definidas principalmente no centro, no leste e no oeste da imagem.

Assim como anteriormente para o algoritmo SVM, notou-se também uma melhora com os dados altimétricos na identificação das áreas destinadas à agricultura nas partes mais planas da bacia, evidenciadas, em sua maior parte, no extremo leste e oeste da imagem, nas cores entre verde claro, verde limão e verde musgo. A savana florestada, na cor verde escuro, apresentou melhoras dentro da área na bacia quando comparada com as classificações anteriores. Nesse cenário, ela é evidenciada, em sua maior parte, perto da área alagada pela hidrelétrica ao norte e quase em todo o seu entorno, bem como em alguns fragmentos no extremo leste.

Os resultados da matriz de confusão (pixels) para o classificador Árvore de Decisão com a imagem utilizando os dados de altimetria apresentados na Tabela 16 demonstram as confusões entre as classes definidas para este estudo.

**Tabela 16** - Matriz de confusão (pixels), erro de omissão e comissão, exatidão global e índice *kappa* para a Árvore de Decisão (DT) com altimetria

<b>Linhas e Colunas</b>	<b>Água</b>	<b>Savana Florestada</b>	<b>Água Turva</b>	<b>Savana Rasteira Antropizada</b>	<b>Solo Exposto</b>	<b>Agricultura Inicial</b>	<b>Agricultura Média</b>	<b>Agricultura Avançada</b>	<b>Savana Parque</b>	<b>Savana Arborizada</b>	<b>Total</b>
<b>Água</b>	54480	0	0	0	0	0	0	0	0	0	54480
<b>Savana Florestada</b>	0	11212	0	10	0	0	0	347	0	895	12464
<b>Água Turva</b>	0	0	7728	0	0	0	0	0	0	0	7728
<b>Savana Rasteira Antropizada</b>	0	0	0	34178	0	23	0	0	0	3329	37530
<b>Solo Exposto</b>	0	0	0	94	12230	0	0	0	0	0	12324
<b>Agricultura Inicial</b>	0	0	0	110	1666	18949	0	0	1449	0	22174

<b>Agricultura Média</b>	0	95	0	0	0	0	15135	0	0	23	15253
<b>Agricultura Avançada</b>	0	8366	0	0	0	0	4	8307	0	463	17140
<b>Savana Parque</b>	0	0	0	199	0	30	0	1	42071	3089	45390
<b>Savana Arborizada</b>	0	305	0	0	0	0	0	0	2617	18488	21410
<b>Total</b>	54480	19978	7728	34591	13896	19002	15139	8655	46137	26287	245893
<b>Erro de Comissão</b>	0%	43,8%	0%	1,19%	11,9%	0,2%	0,02%	4%	8,8%	29,6%	
<b>Erro de Omissão</b>	0%	10%	0%	8,9%	0,7%	14,5%	0,7%	51,5%	7,3%	13,6%	
<b>Exatidão Global</b>	0,91										
<b>Índice Kappa</b>	0,89										

Fonte: Orfeo Toolbox (OTB, 2020).

As classes que tiveram as maiores confusões para os dados citados foram também a agricultura inicial, com um total de 22.174 pixels, dos quais 1.666 foram confundidos com solo exposto e 1.449 com savana parque; a savana rasteira antropizada, com total de 37.530 pixels, em que 3.329 foram confundidos com savana arborizada; a agricultura avançada, com total de 17.140 pixels, dentre os quais 8.366 foram confundidos com savana florestada; a savana parque, com o total de 45.390 pixels, dos quais 3.089 confundiram-se com savana arborizada; por fim, dos 21.410 pixels da savana arborizada, 2.617 foram para a savana parque.

As classes que continuaram com nenhuma ou pouca confusão na matriz foram: água, água turva, solo exposto e agricultura média. Estas confusões entre as classes evidenciadas na matriz se deram, possivelmente, por causa dos limites estabelecidos entre a mesma e sua vizinhança e por possuírem valores de refletâncias e números digitais constituídos para cada pixel da imagem próximos.

A partir do resultado da matriz de confusão foi gerado o índice *kappa*. Apesar das confusões apresentadas entre algumas classes, este classificador mostrou ter bons resultados na hora de classificar uma imagem, obtendo um índice *kappa* de 0,89, que demonstra estar dentro do nível de concordância excelente, conforme Tabela 2. Ademais, se encontra entre os três algoritmos abordados para a imagem com os dados altimétricos, conseguindo um resultado um pouco inferior ao algoritmo de SVM para este cubo de dados, mas obtendo um resultado melhor em referência ao classificador de Floresta Aleatória (RF), também abordado neste estudo.

O índice de exatidão global, por seu turno, mostrou resultado satisfatório de 0,91, configurando-se acima do resultado apresentado pelo índice *kappa* para esta classificação de imagem e com este classificador, e acima também da exatidão global do mesmo classificador, mas para a imagem sem altimetria.

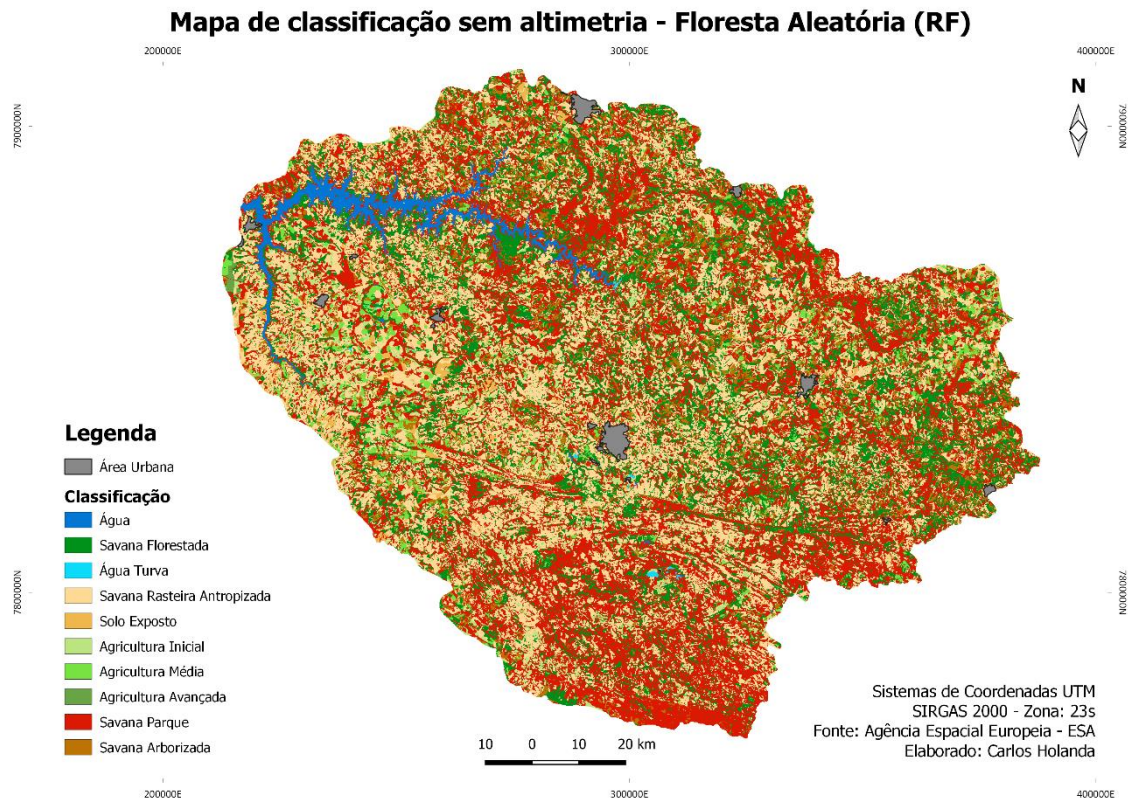
De outra via, os resultados para o erro de omissão mostraram os maiores erros para a classe agricultura avançada, com resultado de 51,5% do total, reafirmando o que a matriz de confusão apresentou na Tabela 13. Para os resultados do erro de comissão, os maiores dados apresentaram-se sobre as classes savana florestada, com o resultado de 43,8%; e savana arborizada, com 29,6%.

Por fim, quanto ao tempo gasto para a realização do processamento dos dados da classificação e pós-classificação desta imagem para este classificador, somam-se:

a) Estatística da imagem; b) Treinamento da imagem com o classificador, em que também são gerados o modelo e a matriz de confusão; c) Classificação de imagem; e d) Pós-classificação da imagem, que foi de, aproximadamente, 40 minutos seguidos, mesmo utilizando-se a imagem com os dados complementares de altimetria.

**5.1.5 Imagem classificada sem altimetria para o algoritmo de Floresta Aleatória (RF)**

Para a observação do algoritmo Floresta Aleatória (RF), a Figura 20 traz o mapa de classificação do uso e cobertura da terra sem altimetria. Neste tipo de cobertura, o algoritmo de RF mostrou ser, dentre os classificadores apoiados em aprendizado de máquina testados nesta pesquisa, aquele que obteve o menor resultado em relação aos algoritmos abordados para a área de estudo com a imagem sem os dados altimétricos e derivações.



**Figura 20 - Mapa de classificação sem altimetria – Floresta Aleatória (RF)**  
Fonte: Agência Espacial Europeia - ESA.

Quando considerada a classificação sem altimetria com RF, o classificador apresentou visualmente uma maior confusão entre as classes dentro da área estudada. Apesar disso conseguiu abordar toda a região sul e norte da bacia hidrográfica mencionadas anteriormente para os algoritmos SVM e DT como savana parque e savana arborizada. Pode-se evidenciar também que estas classes estão distribuídas em quase todas as áreas da bacia hidrográfica para este classificador. Nesses achados destaca-se que, no extremo leste e centro da imagem, há uma maior quantidade de área da savana arborizada e savana parque em relação à classificação do SVM e do DT sem altimetria.

No que tange à classe água, o classificador continuou mostrando toda a área alagada pela barragem de Nova Ponte, como também as pequenas áreas com a classe água turva. Mais uma vez, conforme pode ser observado na imagem classificada e nos dados contidos na matriz de confusão, estas duas classes não sofreram muita alteração.

De outro modo, o classificador não conseguiu identificar, como os anteriores, a grande quantidade de regiões com savana rasteira antropizada, na cor bege claro, e solo exposto, na cor bege escuro, em relação aos classificadores anteriores para a área de estudo. Neste contexto em específico, nota-se que estas classes também estão espalhadas em quase todas as áreas da bacia, mas que, para os classificadores anteriores, há uma maior ênfase para o centro e o oeste da mesma.

Nesta oportunidade, não foi possível notar, entretanto, as grandes quantidades de áreas de agricultura evidenciadas e bem definidas em sua maior parte nos classificadores SVM e DT sem altimetria no extremo leste e oeste da imagem. Logo, não se conseguiu abordar, como um todo, as áreas de agricultura observadas nos algoritmos anteriores também sem altimetria e observadas na imagem do Google Earth, bem como no histórico da área no agronegócio.

A classe savana florestada, na cor verde escuro, apresentou visualmente muitas mudanças dentro da área na bacia quando comparada com as classificações anteriores sem altimetria, podendo ainda ser evidenciada, perto da área alagada pela hidrelétrica ao norte e em alguns fragmentos no extremo leste, mas com uma menor quantidade, se comparada aos anteriores.



Os resultados da matriz de confusão (pixels) para o classificador Floresta Aleatória para a imagem sem altimetria são apresentados na Tabela 17 e demonstram as confusões entre as classes definidas para este estudo.

**Tabela 17** - Matriz de confusão (pixels), erro de omissão e comissão, exatidão global e índice *kappa* para a Floresta Aleatória (RF) sem altimetria

<b>Linhas e Colunas</b>	<b>Água</b>	<b>Savana Florestada</b>	<b>Água Turva</b>	<b>Savana Rasteira Antropizada</b>	<b>Solo Exposto</b>	<b>Agricultura Inicial</b>	<b>Agricultura Média</b>	<b>Agricultura Avançada</b>	<b>Savana Parque</b>	<b>Savana Arborizada</b>	<b>Total</b>
<b>Água</b>	54479	1	0	0	0	0	0	0	0	0	54480
<b>Savana Florestada</b>	0	11952	0	0	0	0	276	18	23	195	12464
<b>Água Turva</b>	0	0	7728	0	0	0	0	0	0	0	7728
<b>Savana Rasteira Antropizada</b>	0	101	0	31018	91	233	12	0	4642	1433	37530
<b>Solo Exposto</b>	0	0	0	254	12070	0	0	0	0	0	12324
<b>Agricultura Inicial</b>	0	448	0	10876	220	7707	1	0	2922	0	22174

<b>Agricultura Média</b>	0	223	0	4	0	0	13252	0	0	1774	15253
<b>Agricultura Avançada</b>	0	13207	0	0	0	0	0	3924	1	8	17140
<b>Savana Parque</b>	0	50	0	1629	0	1232	0	0	42449	30	45390
<b>Savana Arborizada</b>	0	2514	0	617	0	29	16	0	12411	5823	21410
<b>Total</b>	54479	28496	7728	44398	12381	9201	13557	3942	62448	9263	245893
<b>Erro de Comissão</b>	0%	58%	0%	30,1%	2,5%	16,2%	2,2%	0,4%	32%	37,1%	
<b>Erro de Omissão</b>	0%	4,1%	0%	17,3%	2%	65,2%	13,1%	77,1%	6,4%	72,8%	
<b>Exatidão Global</b>	0,77										
<b>Índice Kappa</b>	0,74										

Fonte: Orfeo Toolbox (OTB, 2020).

Neste caso, os achados da pesquisa demonstram que as classes que tiveram as maiores confusões para os dados citados foram também a agricultura inicial, com um total de 22.174 pixels, dos quais 10.876 foram confundidos com savana rasteira antropizada e 2.922 com savana parque; a agricultura média, com total de 15.253 pixels, dos quais 1.774 foram para savana arborizada; a agricultura avançada, com total de 17.140 pixels, dentre os quais 13.207 foram confundidos com savana florestada; a savana rasteira antropizada, que dos 37.530 pixels totais, 4.642 foram confundidos com savana parque e 1.433 foram para a savana arborizada; a savana parque, com total de 45.390 pixels, em que 1.629 foram confundidos com savana rasteira antropizada e 1.232 para agricultura inicial; e a savana arborizada, com total de 21.410 pixels, dos quais 2.514 foram confundidos com savana florestada e 12.411 para savana parque. As classes que continuaram com nenhuma ou pouca confusão na matriz foram: água, água turva e solo exposto.

Estas confusões entre as classes evidenciadas na matriz ocorreram, possivelmente, por causa dos limites estabelecidos entre a mesma e sua vizinhança, e por possuírem valores de refletâncias e números digitais constituídos para cada pixel da imagem próximos. Salienta-se, ainda, que este classificador apresentou as maiores confusões entre as classes para todos os algoritmos testados, utilizando as imagens sem os dados altimétricos e também com os dados.

Nesse íterim, a partir do resultado da matriz de confusão foi gerado o índice *kappa*. Apesar das confusões observadas entre algumas classes e sendo o classificador aquele que apresentou os piores resultados entre os três abordados, o RF mostrou ter bons resultados na hora de classificar uma imagem, obtendo um índice *kappa* de 0,74, e demonstrando que está dentro do nível de concordância muito bom, conforme a Tabela 2.

Por seu turno, o índice de exatidão global também evidenciou resultado satisfatório para a classificação de imagens orbitais, alcançando 0,77. Cumpre lembrar que, como apresentado antes, este valor sempre será maior que o índice *kappa*, mas, para este classificador e com este conjunto específico de dados para a imagem.

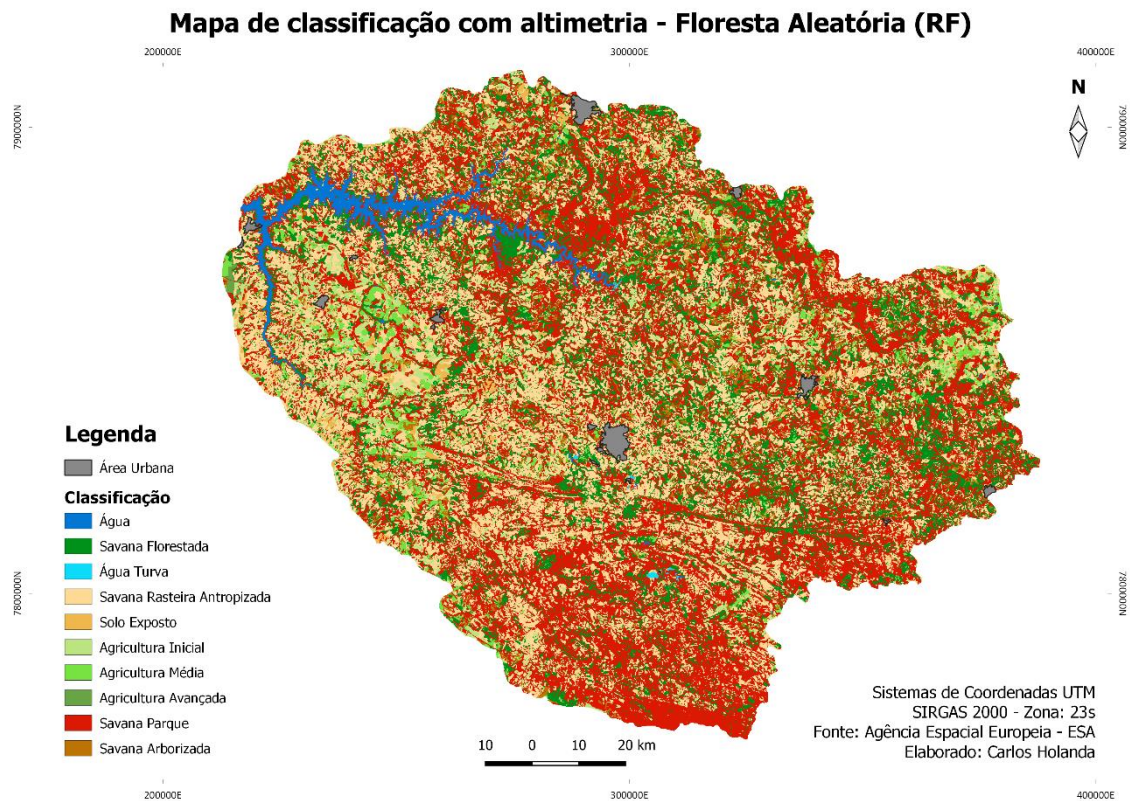
De modo coerente, os resultados para o erro de omissão mostraram os maiores erros para as classes como agricultura inicial, com resultado de 65,2%; agricultura

avançada, com 77,1%; e savana parque, com 72,8%. Para os resultados do erro de comissão, pontua-se que ocorreram os maiores dados para as classes savana florestada, com resultado de 58%; savana rasteira antropizada, com 30,1%; savana parque, com 32%; e savana arborizada, com 37,1% do total.

Por fim, quanto ao tempo gasto para a realização do processamento dos dados da classificação e pós-classificação desta imagem para este classificador, somam-se: a) Estatística da imagem; b) Treinamento da imagem com o classificador, em que também são gerados o modelo e a matriz de confusão; c) Classificação de imagem; e d) Pós-classificação da imagem, que foi de aproximadamente 50 minutos seguidos.

#### **5.1.6 Imagem classificada com altimetria para o algoritmo de Floresta Aleatória (RF)**

A Figura 21 traz o mapa de classificação do uso e cobertura da terra com altimetria para o algoritmo Floresta Aleatória (RF). O algoritmo de RF com os dados altimétricos mostrou ser, dentre os classificadores apoiados em aprendizado de máquina testados nesta pesquisa, aquele que obteve o menor resultado em relação aos algoritmos abordados para a área de estudo, assim evidenciando resultados melhores apenas para o mesmo classificador, mas com relação à imagem sem os dados de altimetria.



**Figura 21** - Mapa de classificação com altimetria – Floresta Aleatória (RF)  
Fonte: Agência Espacial Europeia - ESA.

Tem-se, a partir da observação da Figura 21 que, assim como a imagem anterior, o classificador RF mesmo com os dados altimétricos obteve também visualmente uma maior confusão na distribuição das classes como na classificação sem altimetria para o RF em relação aos algoritmos anteriores. Apesar da utilização dos dados altimétricos, o algoritmo não conseguiu definir claramente as partes mais elevadas e planas da área de estudo como visto nos classificadores SVM e DT com estes dados complementares. Apesar de apresentar esta maior distribuição das classes savana arborizada e savana parque em praticamente toda a área, elas ainda são visualizadas, em sua maior parte, no extremo sul da imagem, no extremo norte e nordeste.

Em detrimento desta maior confusão, pode-se observar, também, no extremo leste e centro estas classes como na classificação sem altimetria (RF), que há uma maior quantidade em área visualmente da savana arborizada e savana parque em relação à classificação do SVM e do DT com altimetria.

Para a classe água, o classificador seguiu mostrando toda a área alagada pela barragem de Nova Ponte, como também as pequenas áreas com a classe água turva.

Seguindo o mesmo padrão encontrado na imagem anterior sem altimetria (RF), o classificador mesmo com os dados altimétricos, não conseguiu definir de forma clara a grande quantidade de regiões com savana rasteira antropizada (cor bege claro) e com solo exposto (cor bege escuro), dentro da área de estudo em relação aos algoritmos SVM e DT com os dados altimétricos. Apesar da menor definição destas classes na área de estudo, nota-se uma maior ênfase no centro e no oeste da imagem.

Entretanto, não se notou grandes quantidades de áreas de agricultura evidenciadas e bem definidas como nos classificadores SVM e DT com altimetria no extremo leste e oeste da imagem. Todavia, foram alcançadas melhoras em relação ao mesmo classificador, mas sem os dados altimétricos.

Nesse sentido, a classe savana florestada (cor verde escuro) também apresentou visualmente muitas mudanças dentro da área da bacia quando comparada com as classificações dos algoritmos SVM e DT com altimetria, mas, no entanto, apresentou melhores resultados em relação ao mesmo classificador sem os dados altimétricos. Esta classe pode, ainda, ser evidenciada, em sua maior parte, perto da área alagada pela hidrelétrica ao norte e em alguns fragmentos no extremo leste.

Os resultados da matriz de confusão (pixels) para o classificador Floresta Aleatória para a imagem com altimetria são apresentados na Tabela 18 e demonstram as confusões entre as classes definidas para este estudo.

**Tabela 18** - Matriz de confusão (pixels), erro de omissão e comissão, exatidão global e índice *kappa* para a Floresta Aleatória (RF) com altimetria

<b>Linhas e Colunas</b>	<b>Água</b>	<b>Savana Florestada</b>	<b>Água Turva</b>	<b>Savana Rasteira Antropizada</b>	<b>Solo Exposto</b>	<b>Agricultura Inicial</b>	<b>Agricultura Média</b>	<b>Agricultura Avançada</b>	<b>Savana Parque</b>	<b>Savana Arborizada</b>	<b>Total</b>
<b>Água</b>	54480	0	0	0	0	0	0	0	0	0	54480
<b>Savana Florestada</b>	0	12278	0	22	0	0	0	0	135	29	12464
<b>Água Turva</b>	0	0	7728	0	0	0	0	0	0	0	7728
<b>Savana Rasteira Antropizada</b>	0	0	0	34416	0	50	0	0	2992	72	37530
<b>Solo Exposto</b>	0	0	0	142	12182	0	0	0	0	0	12324
<b>Agricultura Inicial</b>	0	0	0	9548	9	10646	428	0	1543	0	22174



<b>Agricultura Média</b>	0	552	0	0	0	0	14478	0	0	223	15253
<b>Agricultura Avançada</b>	0	13817	0	3	0	0	0	3317	1	2	17140
<b>Savana Parque</b>	0	0	0	1113	0	156	0	0	44121	0	45390
<b>Savana Arborizada</b>	0	2419	0	0	0	0	0	0	17974	1017	21410
<b>Total</b>	54480	29066	7728	45244	12191	10852	14906	3317	66766	1343	245893
<b>Erro de Comissão</b>	0%	57,7%	0%	23,9%	0%	1,8%	2,8%	0%	33,9%	24,2%	
<b>Erro de Omissão</b>	0%	1,4%	0%	8,2%	1,1%	51,9%	5%	80,6%	2,7%	95,2%	
<b>Exatidão Global</b>	0,79										
<b>Índice Kappa</b>	0,76										

Fonte: Orfeo Toolbox (OTB, 2020).

Nesse cenário, as classes que tiveram as maiores confusões para os dados citados foram também a agricultura inicial, com um total de 22.174 pixels, dos quais 9.548 foram confundidos com savana rasteira antropizada e 1.543 foram para a savana parque; a agricultura avançada, com total de 17.140 pixels, em que 13.817 foram confundidos com savana florestada; a savana parque, com total de 45.390 pixels, dentre os quais 1.113 foram confundidos com savana rasteira antropizada; e a savana arborizada, com total de 21.410 pixels, dos quais 2.419 foram confundidos com savana florestada e 17.974 foram para a savana parque.

As classes que continuaram com nenhuma ou pouca confusão na matriz foram: água, água turva e solo exposto. Estas confusões entre as classes evidenciadas na matriz ocorreram, possivelmente, por causa dos limites estabelecidos entre a mesma e sua vizinhança e por possuírem valores de refletâncias e números digitais constituídos para cada pixel da imagem próximos. Salieta-se, ainda, que este classificador apresentou as maiores confusões entre as classes para os algoritmos testados na imagem composta com os dados altimétricos, mas obteve melhor resultado em relação com a classificação do mesmo algoritmo sem os dados de altimetria.

Nesta etapa do estudo, por meio do resultado da matriz de confusão foi gerado o índice *kappa*. Apesar das confusões observadas entre algumas classes e sendo o classificador aquele que possui os piores resultados entre os três abordados para a imagem com as informações altimétricas, este mostrou ter bons resultados na hora de classificar uma imagem, obtendo um índice *kappa* de 0,76, demonstrando, dessa maneira, que está dentro do nível de concordância muito bom, conforme a Tabela 2.

O índice de exatidão global também apresentou resultado satisfatório para a classificação de imagens orbitais, alcançando 0,79. Como apresentado antes, este valor sempre será maior que o índice *kappa*, mas para este classificador e com este conjunto de dados para a imagem.

Os resultados para o erro de omissão mostraram os maiores erros para as classes como agricultura inicial, com resultado de 51%; agricultura avançada, com 80,6%; e savana arborizada, com 95,2% do total. Para os resultados do erro de comissão, os maiores achados se deram para as classes: savana florestada, com

resultado de 57,7%; savana rasteira antropizada, com 23,9%; savana parque, com 33,9%; e savana arborizada, com 24,2%.

Por fim, quanto ao tempo gasto para a realização do processamento dos dados da classificação e pós-classificação desta imagem para este classificador, somam-se: a) Estatística da imagem; b) Treinamento da imagem com o classificador, em que também são gerados o modelo e a matriz de confusão; c) Classificação de imagem; e d) Pós-classificação da imagem, que foi de aproximadamente 50 minutos seguidos também, mesmo utilizando a imagem com os dados complementares de altimetria.

### **5.1.7 Comparação dos resultados das matrizes de confusão dos algoritmos SVM, DT e RF com e sem os dados altimétricos**

A comparação dos resultados encontrados nas matrizes de confusão com os melhores e piores desempenhos tanto para as imagens classificadas com os dados altimétricos (modelos digitais de elevação), tanto com as imagens classificadas sem estas informações adicionais (sem os modelos) estão apresentadas na Tabela 19. Os resultados na diagonal na Tabela 19 representados na cor verde mostram os algoritmos que obtiveram as melhores concordâncias nas matrizes de confusão sem e com os dados altimétricos para cada classe abordada. Os algoritmos representados fora da diagonal na cor laranja foram os piores resultados encontrados nas matrizes de confusão para cada classe, ou seja, o maior nível de confusão entre as classes. Os dados em branco foram os resultados que obtiveram nenhuma confusão nas matrizes.

**Tabela 19** – Quadro comparativo dos resultados das matrizes de confusão dos algoritmos SVM, DT e RF com e sem os dados altimétricos. O algoritmo que apresentou o melhor resultado para cada classe está marcado em verde, enquanto o que apresentou a maior confusão entre as classes está em laranja.

Linhas e Colunas	Água	Savana Florestada	Água Turva	Savana Rasteira Antropizada	Solo Exposto	Agricultura Inicial	Agricultura Média	Agricultura Avançada	Savana Parque	Savana Arborizada
Água	SVM sem/com modelo DT sem/com modelo RF sem modelo	RF sem modelo								
Savana Florestada		RF com modelo		RF com modelo			RF sem modelo	DT sem modelo	RF com modelo	DT com modelo
Água Turva			SVM sem/com modelo DT sem/com modelo RF sem modelo							
Savana Rasteira		RF sem modelo		RF com modelo	DT sem modelo	RF sem modelo	SVM sem modelo	DT sem modelo	RF sem modelo	DT sem modelo

<b>Antropizada</b>										
<b>Solo Exposto</b>				RF sem modelo	RF com modelo					
<b>Agricultura Inicial</b>		RF sem modelo		RF sem modelo	SVM com modelo	DT sem modelo	RF com modelo		RF sem modelo	DT sem modelo
<b>Agricultura Média</b>		RF com modelo		DT sem modelo			DT com modelo	SVM com modelo		RF sem modelo
<b>Agricultura Avançada</b>		RF com modelo		RF com modelo			DT sem modelo	SVM com modelo	RF sem/com modelo	DT sem modelo
<b>Savana Parque</b>		RF sem modelo		SVM sem modelo	DT sem modelo	RF sem modelo	SVM sem modelo	DT com modelo	RF com modelo	DT sem modelo
<b>Savana Arborizada</b>		RF sem modelo		RF com modelo	DT sem modelo	RF com modelo	RF com modelo	DT sem modelo	RF com modelo	DT com modelo

Fonte: Orfeo Toolbox (OTB, 2020).

Como mencionado anteriormente, a classificação que obteve os melhores desempenhos entre os três classificadores foi a do algoritmo de SVM, para os dois conjuntos de imagens (Tabela 20). Apesar do bom desempenho deste algoritmo na classificação de imagens de satélite, o classificador obteve o tempo de processamento muito superior para gerar a classificação final entre os três abordados e para os dois cubos de imagens. O algoritmo de DT obteve resultados igualmente satisfatórios, mas com tempo gasto de processamento significativamente menor em relação ao de SVM para os dois cubos de dados. De acordo com o objetivo que o pesquisador queira alcançar, levando-se em conta o tamanho em extensão de área de estudo que ele precise classificar e o tempo gasto em processamento, este classificador é mais adequado que o anterior para este cenário como mostra a Tabela 20.

**Tabela 20** - Quadro comparativo final dos resultados *kappa* e tempo de processamento obtidos nos classificadores abordados e no mapa da ESRI

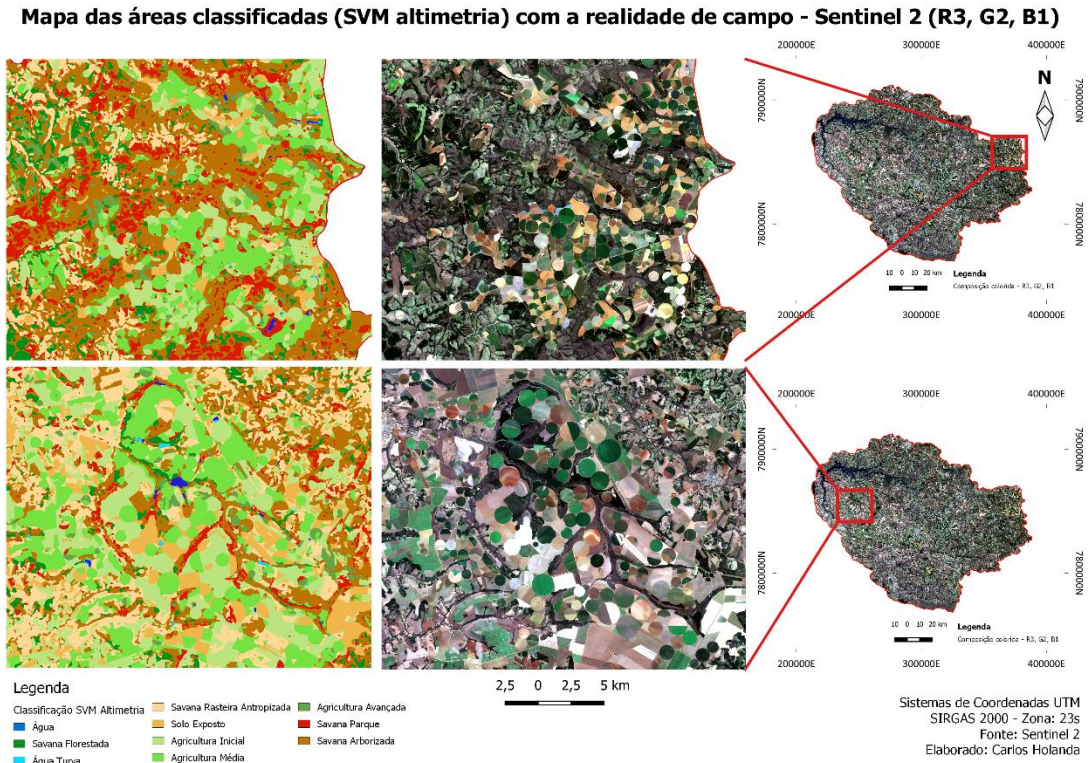
<b>Resultados <i>kappa</i> e tempo de processamento para os classificadores e mapa ESRI</b>	<b>Algoritmo SVM</b>	<b>Algoritmo DT</b>	<b>Algoritmo RF</b>	<b>Mapa ESRI</b>
<b>Sentinel 2</b>	0,89 / 48h	0,87 / 40min	0,74 / 50min	0,78
<b>Sentinel 2 + dados altimétricos</b>	<b>0,91</b> / 48h	0,89 / <b>40min</b>	0,76 / 50min	

Fonte: Orfeo Toolbox (OTB, 2020).

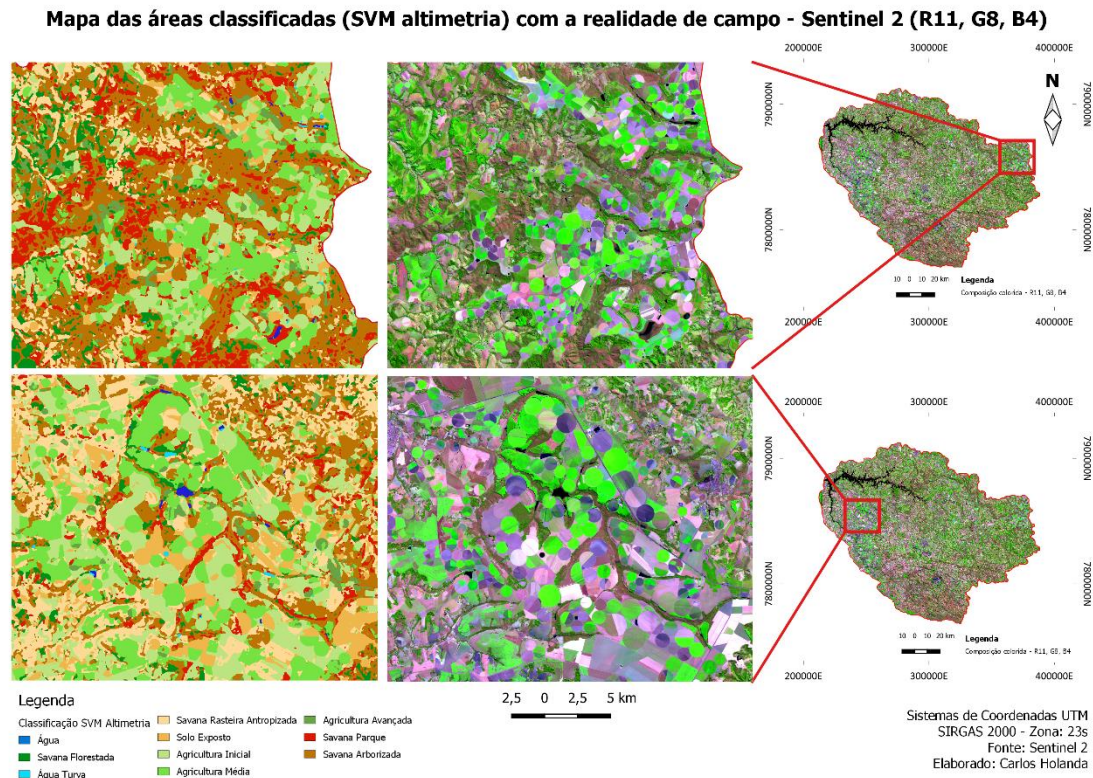
### **5.1.8 Mapa das áreas classificadas (SVM altimetria) com a realidade de campo – Sentinel 2**

A imagem da classificação que obteve os melhores índices estatísticos foi o mapa gerado pelo algoritmo SVM com os dados altimétricos como mencionado anteriormente. Assim sendo, este mapa da classificação foi comparado com duas composições de bandas da mesma imagem Sentinel 2 mostrando a realidade destas

classes em campo. A primeira imagem (Figura 22), mostram duas partes da área de estudo com as possíveis classes obtidas na classificação e a realidade de campo na composição colorida da imagem Sentinel 2. A segunda imagem (Figura 23), mostram também as classes encontradas pelo algoritmo e a realidade de campo, mas agora na composição falsa cor para a imagem Sentinel 2.



**Figura 22 - Mapa classificação (SVM) com Sentinel-2 composição colorida**  
Fonte: Agência Espacial Europeia - ESA.



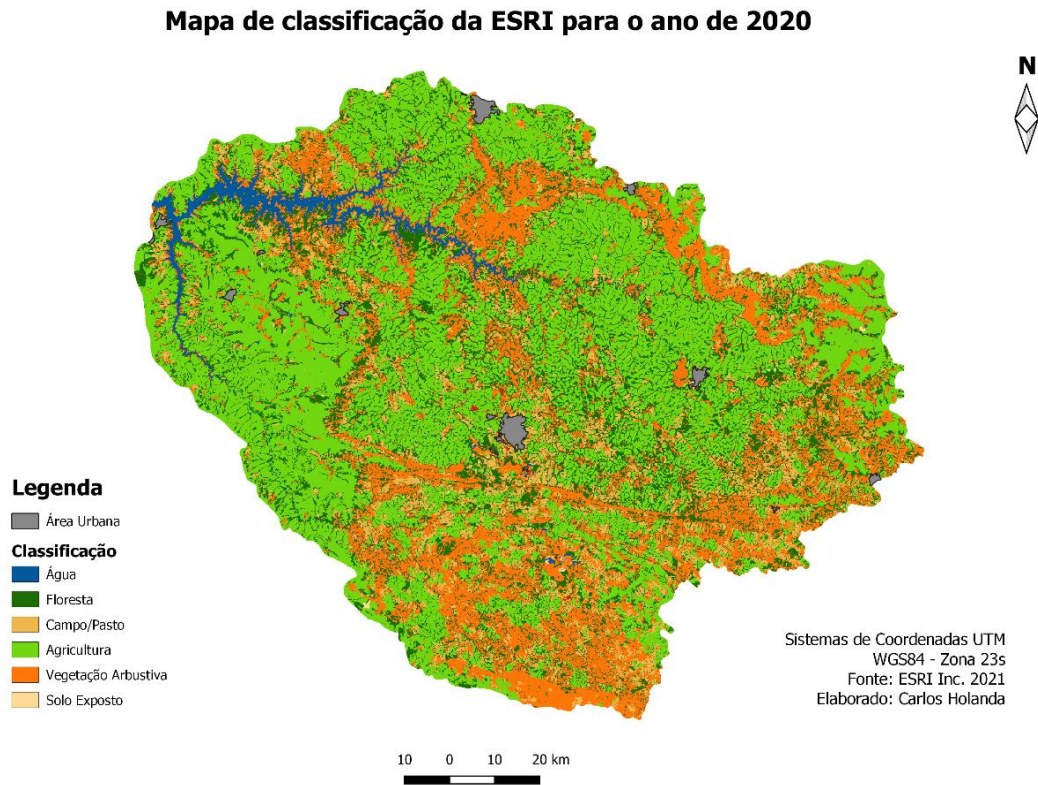
**Figura 23** - Mapa classificação (SVM) com Sentinel-2 composição falsa cor  
Fonte: Agência Espacial Europeia - ESA.

Os resultados tanto para a imagem do Sentinel 2 na composição colorida quanto para a composição em falsa cor demonstraram a qualidade do classificador (SVM com altimetria) em representar as classes de acordo com o que está caracterizado nas imagens de satélite. Estes recortes foram escolhidos por serem áreas de intensa agricultura na região, demonstrando as classificações nos diferentes estágios de crescimento de plantio, pivôs centrais, e as classes como solo exposto na cor bege escuro e savana rasteira antropizada na cor bege claro, que fazem limites com as áreas agricultáveis. Podemos observar também a similaridade das áreas de savana parque em vermelho e savana arborizada na cor marrom com as áreas observadas nas imagens do Sentinel 2.

## 5.2 Mapa de uso e cobertura da terra da ESRI para o ano de 2020



O mapa de classificação do uso e cobertura da terra produzido pela ESRI para o ano de 2020 com o auxílio de imagens do Sentinel 2 pode ser visualizado na Figura 24.



**Figura 24** - Mapa de classificação da ESRI para o ano de 2020  
Fonte: ESRI (2021).

O classificador do mapa da ESRI (Figura 24) conseguiu também abordar, em sua maior parte, toda a região sul da bacia hidrográfica e extremo norte, mas agora classificada como vegetação arbustiva na cor laranja. Isto pode ser visualizado também nas imagens passadas como a união das classes savana parque e savana arbustiva. Visualmente, a ferramenta apresentou uma confusão maior para esta classe no extremo sudeste da imagem em comparação com as imagens anteriores.

Já para a classe água, o classificador mostrou toda a área alagada pela barragem de Nova Ponte, como também as pequenas áreas com a classe água turva, mas agora como uma única classe água.

O classificador não conseguiu identificar, como nas situações anteriores, a grande quantidade de regiões com a classe campo/pasto, na cor bege, classificado

antes como savana rasteira antropizada, e solo exposto, na cor bege claro, classificado como solo exposto também nas imagens passadas, em relação aos classificadores anteriores para a área de estudo. Foi possível notar que as grandes quantidades de áreas de agricultura são agora evidenciadas em quase toda a imagem na cor verde.

Em relação às imagens anteriores e observando a imagem do Google Earth, onde esta classe foi evidenciada em sua maior parte nos extremos oeste e leste, este mapa teve uma maior confusão em relação a sua distribuição para a classe agricultura. A classe floresta, na cor verde escuro, representada antes como savana florestada, apresentou visualmente muitas mudanças dentro da área na bacia quando comparada com as classificações anteriores, podendo ainda ser evidenciada, em sua maior parte, perto da área alagada pela hidrelétrica ao norte, e em alguns fragmentos no extremo leste e sul, mas com uma menor quantidade, se comparada aos anteriores.

Os resultados da matriz de confusão (pixels) para o mapa da ESRI estão apresentados na Tabela 21 e demonstram as confusões entre as classes encontradas para o recorte da área de estudo.

**Tabela 21** - Matriz de confusão (pixels), exatidão global e índice *kappa* para o mapa da ESRI

<b>Linhas e Colunas</b>	<b>Água</b>	<b>Floresta</b>	<b>Campo/Pasto</b>	<b>Agricultura</b>	<b>Vegetação Arbustiva</b>	<b>Solo Exposto</b>	<b>Total</b>
<b>Água</b>	58367	0	0	0	0	0	58367
<b>Floresta</b>	0	12479	0	0	0	0	12479
<b>Campo/Pasto</b>	0	0	22175	9662	4961	11	36809
<b>Agricultura</b>	0	5138	275	48821	303	0	54537
	0	3636	4925	1788	63139	0	73488

<b>Vegetação Arbustiva</b>							
<b>Solo Exposto</b>	0	0	143	12177	0	0	12320
<b>Total</b>	58367	21253	27518	72448	68403	11	248000
<b>Exatidão Global</b>	0,83						
<b>Índice Kappa</b>	0,78						

Fonte: ESRI (2021).

As classes que tiveram as maiores confusões para os dados citados foram o campo/pasto, com um total de 36.809 pixels, dos quais 9.662 foram confundidos com agricultura e 4.961 foram para vegetação arbustiva; a agricultura, com total de 54.537 pixels, em que 5.138 foram confundidos com floresta; a vegetação arbustiva, com total de 73.488 pixels, dentre os quais 3.636 foram confundidos com floresta, 4.925 foram para campo/pasto e 1.788 foram confundidos com agricultura.

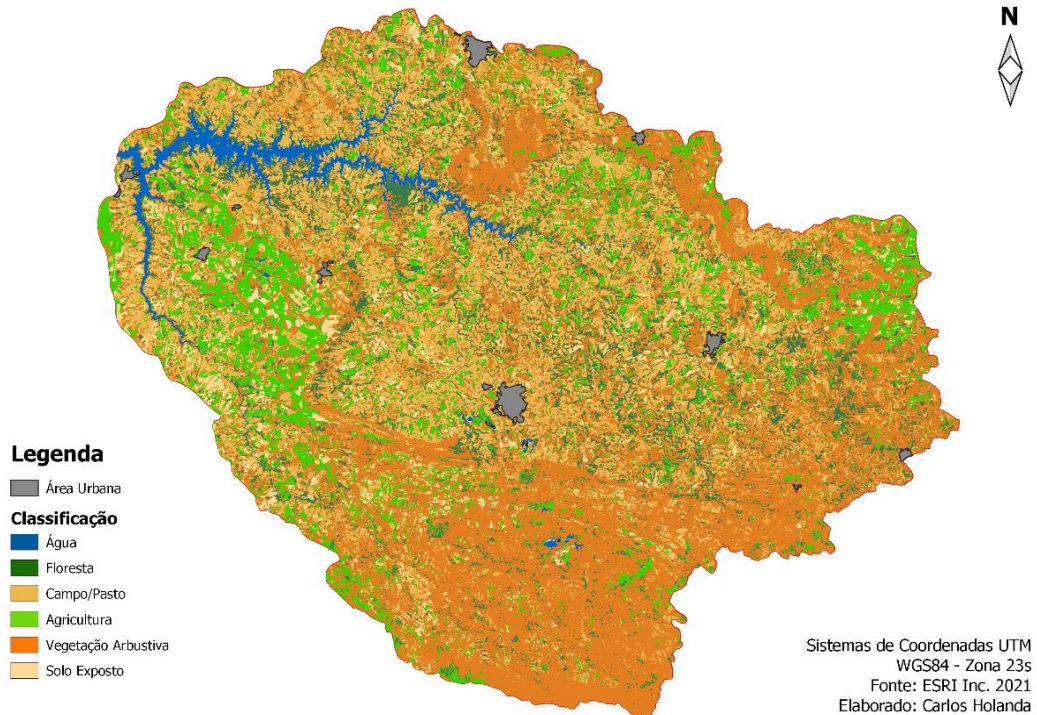
Apesar das confusões observadas entre algumas classes e sendo o classificador aquele que possui os piores resultados entre os três abordados para a imagem com as informações altimétricas, este O algoritmo utilizado para gerar o mapa de ESRI mostrou ter bons resultados na hora de classificar uma imagem, obtendo um índice *kappa* de 0,78, demonstrando, dessa maneira, que está dentro do nível de concordância muito bom, conforme a Tabela 2, apesar de ter ficado abaixo dos índices *kappa* dos algoritmos SVM e DT, acima apenas do RF.

O índice de exatidão global também apresentou resultado satisfatório para a classificação, alcançando 0,83. Como apresentado antes, este valor sempre será maior que o índice *kappa*.

Ainda que os dados estatísticos do mapeamento do algoritmo utilizado pela ESRI foram menores que os apresentados nesta pesquisa, mostrou ser uma ferramenta eficaz e gratuita para a análise do momento atual que se encontra o uso e cobertura da terra do mundo. O mapeamento da classificação do uso e cobertura da terra em escalas regionais e locais, utilizando informações com maior exatidão, como por exemplo o conhecimento prévio da área, revelou ser mais eficiente que um mapeamento em escala global.

Para uma melhor comparação dos mapas gerados por esta pesquisa com o mapa produzido pela ESRI, foi selecionado o mapa do algoritmo SVM com os dados altimétricos que apresentou os melhores índices estatísticos, agregando as classes encontradas neste estudo com as classes apontadas no mapa da ESRI.

**Mapa da classificação (SVM altimetria) com as mesmas classes da ESRI**



**Figura 25** - Mapa de classificação (SVM altimetria) com as classes da ESRI  
Fonte: Agência Espacial Europeia - ESA.

O mapa da classificação SVM com os dados complementares de altimetria, mas agora agregando as classes deste estudo com as classes apresentadas pela ESRI, mostrou uma melhor distribuição das áreas classificadas mesmo juntando as classes em relação ao mapa anterior.

A classe agricultura por exemplo, evidenciada principalmente no extremo oeste e leste da Figura 25 na cor verde claro, apresentou visualmente melhor divisão com as classes campo/pasto na cor bege escuro e solo exposto na cor bege claro, estas observadas em sua maior parte no centro da imagem, em contrapartida com o mapa da ESRI que englobou estas áreas como agricultura em sua maior parte.

Os dois mapas apresentaram relativamente classificações parecidas em relação às classes vegetação arbustiva na cor laranja e floresta na cor verde escuro, apesar da vegetação arbustiva mostrar maior quantidade no extremo sul, leste e norte do mapa da Figura 25 comparado com o mapa da ESRI.

## 6 CONCLUSÃO

A análise e a comparação dos resultados encontrados nas classificações de imagens utilizando algoritmos baseados em aprendizagem de máquina para a área de estudo que está inserida no bioma Cerrado mostraram-se eficientes na identificação das classes encontradas. A partir dessas imagens foram aplicadas técnicas de processamento digital de imagens (PDI) e sensoriamento remoto (SR), aliadas com a aprendizagem de máquina.

A utilização do cubo de dados com a imagem do SRTM e suas derivações, como declividade e rugosidade na classificação do uso e cobertura da terra, mostrou melhores resultados em comparação com as imagens sem esses dados altimétricos. O produto desta classificação, ao se analisar a matriz de confusão e visualizar a imagem, conseguiu distribuir melhor e mais uniformemente as classes na área da bacia hidrográfica, revelando que estas informações adicionais no conjunto das bandas para a formação da imagem final obtiveram melhoras na classificação de imagens de satélite, de acordo com as validações abordadas para testar as mesmas em relação às imagens sem os dados de altimetria e derivações.

Visualmente foi possível identificar que os classificadores conseguiram abordar toda a região sul da bacia hidrográfica, classificada como savana parque e savana arborizada, as quais representam as áreas altimétricas mais elevadas. Em todas as imagens classificadas foi possível notar as áreas destas classes citadas acima, sendo visualizadas em sua maior parte no extremo sul, no extremo norte e nordeste. Pode-se evidenciar também que em todos os algoritmos utilizados a classe água mostrou toda a área alagada pela barragem de Nova Ponte e as pequenas áreas, com a classe água turva, como bacias de rejeitos.

Entre as imagens classificadas foi possível observar também que a grande quantidade de regiões com savana rasteira antropizada, na cor bege claro, e com solo

exposto, na cor bege escuro, com ênfase no centro e oeste dela. Os classificadores conseguiram abordar também as áreas de agricultura nos seus diversos estágios evolutivos, mas identificadas em sua maior parte no extremo leste e oeste da imagem, mostrando predomínio da agricultura. A classe savana florestada, na cor verde escuro, que mostram as áreas que não sofreram alteração pelo homem não apresentaram grandes quantidades dentro da bacia hidrográfica, encontradas em sua maior parte perto da área alagada, ao norte, e alguns fragmentos no extremo leste.

Apesar da diferença nos resultados encontrados nesta pesquisa, tanto para a classificação sem e com os dados complementares de altimetria, como para os diferentes algoritmos, os três classificadores abordados conseguiram ter desempenho satisfatório dentro dos valores aceitáveis de validação para a classificação de imagens de satélite, os quais podem ser observados, principalmente, no índice *kappa*. Como demonstrado antes, para todos os classificadores, a utilização de uma composição de imagens que incrementa seus dados com informações altimétricas para a geração da classificação do uso e cobertura da terra tem melhores resultados na precisão dos conflitos entre as classes e na sua identificação, que podem ser demonstrados nos dados dos validadores empregados.

A classificação que obteve os melhores desempenhos entre todas as classificações produzidas foi a do algoritmo de SVM, tanto para as imagens com os dados altimétricos quanto para a imagem sem estas informações complementares. Este resultado possivelmente pode ser explicado pelo bom desempenho deste classificador em trabalhar um conjunto de dados que possuem diversos atributos, independentemente se tiver poucas amostras de treinamento.

Apesar desse algoritmo ter mostrado bons desempenhos na hora de classificar uma imagem de satélite, ele foi o classificador que obteve também um tempo de processamento muito superior para gerar a classificação final entre os três abordados e para os dois cubos de imagens. Logo, dependendo do objetivo que o pesquisador deseja alcançar e o tempo gasto que ele precisa para gerar suas classificações, caso não possua uma máquina com processador adequado e/ou que seja uma área extensa de estudo, ou várias pequenas áreas, tal classificador não seria o mais apropriado.

O classificador que obteve resultados igualmente satisfatórios, mas menores que o SVM para os dois cubos de dados, foi o algoritmo de DT. Apesar de não conseguir resultados melhores que o SVM, esse algoritmo mostrou ter bom desempenho entre os três abordados para a classificação de imagens de satélite.

Todavia, apesar de esse classificador não mostrar os melhores resultados entre os três abordados, o tempo gasto de processamento para a classificação é muito menor que para o algoritmo de SVM, tanto para a imagem sem quanto para a imagem com altimetria demonstrado antes. Assim, dependendo do objetivo que o pesquisador queira alcançar, levando-se em conta o tempo gasto de processamento e o tamanho em extensão de área de estudo que ele precise classificar, este classificador é mais adequado que o anterior para este cenário.

Por fim, o classificador que obteve os menores resultados entre os três analisados, tanto para a imagem sem os dados altimétricos como para a imagem com os dados, foi o RF, apesar de também ter mostrado desempenho satisfatório observando-se os validadores aplicados para o estudo. Esse algoritmo mostrou, igualmente, melhoras na classificação quando manuseado com a imagem com o conjunto de dados complementares de altimetria. O tempo gasto de processamento para este classificador foi muito menor também que o do SVM para todo o processo de classificação, mas obteve tempo parecido com o algoritmo de DT para as duas imagens.

Estes resultados podem ser utilizados também em trabalhos futuros integrados com outros índices ópticos ao conjunto de dados como: a) o Índice de Vegetação da Diferença Normalizada (NDVI); b) o Índice de Vegetação Ajustado para os Efeitos do Solo (SAVI); e c) o Índice de Umidade da Diferença Normalizada (NDWI), mencionados anteriormente como alternativa para aumentar a precisão dos resultados da classificação final.

Vale salientar também que as limitações encontradas nesta pesquisa foram a não coleta de pontos de controle e levantamento de dados em campo devido principalmente ao isolamento social provocado pela pandemia que de certa forma comprometeram a validação dos resultados da classificação das imagens. Apesar disso, os dados utilizados para validar e treinar os algoritmos de classificação como mencionado anteriormente foram retirados das imagens do *Google Earth* para o mês

e o ano de estudo e as imagens do Sentinel 2 em diversas composições de bandas. As recomendações levantadas para possíveis trabalhos futuros seriam de fazer as classificações e o treinamento desses algoritmos utilizando estes dados coletados em campo e ver se existem melhoras nos resultados encontrados pelos validadores estatísticos abordados. Outro ponto levantado para estudos futuros seria o de fazer a classificação desses algoritmos em diferentes imagens de satélites com resoluções espaciais (dimensão de pixel) distintas, mais refinadas e mais grosseiras, avaliando se os resultados de validação estatísticos pioram ou melhoram.

Assim sendo, esta pesquisa conseguiu trazer dados novos e relevantes na comparação e análise dos algoritmos de classificação de imagens de satélites para a área de estudo, apresentando, portanto, informações detalhadas sobre o desempenho destes algoritmos para que estudantes e pesquisadores da área de sensoriamento remoto possam utilizar estas informações para melhor direcionar suas pesquisas, dentro de suas realidades, quando forem classificar suas imagens de satélites.

Dessa forma, o trabalho, além de servir como um norte para esses estudantes e pesquisadores, gerou uma importante base de dados do uso e cobertura da terra para a área de estudo, com informações de três algoritmos diferentes, e que os dados gerados de uso e cobertura da terra podem ser utilizados no futuro por agências públicas e privadas em tomadas de decisão, como a CEMIG, que é a administradora da barragem e todo o entorno da bacia hidrográfica. Além disso, serviu também para compor a base de dados de um projeto de pesquisa do qual este pesquisador faz parte, alimentando esta base com informações mais atuais que servirão para a sequência de estudos na área para futuros pesquisadores.



## REFERÊNCIAS

- ADAM E., *et al.* **Land-use/cover classification in a heterogeneous coastal landscape using RapidEye imagery: evaluating the performance of random forest and support vector machines classifiers.** International Journal of Remote Sensing, 35(10), pp. 3440-3458, doi: 10.1080/01431161.2014.903435. 2014.
- AMARAL M. B., RIOS S. A. **Geoprocessamento: mapeamento do uso e ocupação do solo no alto curso do Rio Piedade.** Revista de geografia física, 2:1-8. 2012.
- AKAR, O. *et al.* **Classification of multispectral images using Random Forest algorithm.** Journal of Geodesy and Geoinformation, UCTEA Chamber of Surveying and Cadastre Engineers, Turkey. Vol.1. Iss.2. pp. 105-112. November, 2012.
- AKAR, Ö.; GUNGOR, O. **Integrating multiple texture methods and NDVI to the Random Forest classification algorithm to detect tea and hazelnut plantation areas in northeast Turkey.** Int. J. Remote Sens. 2015, 36, 442–464.
- AHLERT, S. *et al.* **Mapeamento do uso e cobertura do solo do município de Caxias do Sul (RS) através de imagens do satélite CBERS.** Anais XIII Simpósio Brasileiro de Sensoriamento Remoto, Florianópolis, Brasil, 21-26 abril 2007, INPE, p. 1103-1107.
- BERNARDES, T. **Caracterização do ambiente agrícola do Complexo Serra Negra por meio de sensoriamento remoto e sistemas de informação geográfica.** Dissertação (Mestrado). Universidade Federal de Lavras, Minas Gerais.119p. 2006.
- BLASCHKE, T. **Object based image analysis for remote sensing.** ISPRS Journal of Photogrammetry and Remote Sensing, v.65, p.2–16, 2010.
- BONESSO, D. **Estimação dos Parâmetros do Kernel em um Classificador SVM na Classificação de Imagens Hiperespectrias em uma Abordagem Multiclasse.** Dissertação apresentada na UFRGS, Centro Estadual de Pesquisas em Sensoriamento Remoto e Meteorologia, PPGSR, Porto Alegre, 2013.
- BOYD D. S., DANSON F. M. **Satellite remote sensing of forest resources: three decades of research development.** Prog Phys Geog 29: 1–26, 2005.
- BRANDÃO, M. Cerrado. In: **Lista vermelha das espécies ameaçadas de extinção da flora de Minas Gerais** (Orgs. Mendonça & Lins). 2000. p.55 – 63.
- BRASIL. Ministério do Meio Ambiente – MMA. **Avaliação e identificação de áreas e ações prioritárias para a conservação, utilização sustentável e repartição dos benefícios da biodiversidade nos biomas brasileiros.** Brasília, 2002.
- BRASIL. Ministério do Meio Ambiente. **Plano de ação para prevenção e controle do desmatamento e das queimadas: Cerrado.** Brasília: MMA. 200 p. 2011.

BREIMAM, L. **Random Forests**. Kluwer Academic Publishers. Manufactured in The Netherlands. Machine Learning, 45, 5–32, 2001.

BRISCO, B.; BROWN, R.J.; MANORE, M.J. **Early season crop discrimination with combined SAR and TM data**. Can. J. Remote Sens 1989, 15, 44–54.

BRITES, R.S. **Verificação de exatidão em classificação de imagens digitais orbitais: efeito de diferentes estratégias de amostragem e avaliação de índices de exatidão**. Viçosa, 1996, 101 p. Tese (Doutorado em Ciência Florestal) - Universidade Federal de Viçosa.

BRODLEY, C.E., UTOFF, P.E. **Multivariate decision trees**. Mach Learn 19, 45–77, 1993.

CAVALCANTI, L. C. A. M. **Detecção de elementos antrópicos em imagens aéreas da floresta amazônica**. Dissertação (Mestrado em Informática) - Universidade Federal do Amazonas. 94 f.: il. 2016

CHUVIECCO, E. **Fundamentos de Teledetección Espacial**. Madri, España: 2 ed. Ediciones Rialp, 449p. 1995.

COHEN, J. **A coefficient of agreement for nominal scales**. Educational and Psychological Measurement. Nova York, v. 20, n.1, p. 37-46, 1960.

CONGALTON, R. G. **A review of assessing the accuracy of classifications of remotely sensed data**. Remote Sensing of Environment, v. 49 n. 12, p. 1671-1678, 1991.

CONGALTON, R. G.; GREEN, K. **Assessing the accuracy of remotely sensed data: principles and practices**. Lewis Publishers, Boca Raton, FL, 1999.

COUTINHO, A. C. **Segmentação e classificação de imagens landsat-TM para o mapeamento dos usos da terra na região de Campinas, SP**. Universidade de São Paulo, p. 2, 1997.

DU, P.J.; SAMAT, A.; WASKE, B.; LIU, S.C.; LI, Z.H. **Random Forest and Rotation Forest for fully polarized SAR image classification using polarimetric and spatial features**. ISPRS J. Photogramm. Remote Sens. 2015, 105, 38–53.

EITEN, G. **Formas fisionômicas do Cerrado**. Revista Brasileira de Botânica, São Paulo, v. 2, n. 2, p. 139-148, 1979.

ESA. **European Spatial Agency**. 2020. Disponível em: <https://sentinel.esa.int/web/sentinel/missions>. Acesso em: 12 de setembro de 2020.

FERREIRA, E.; DANTAS, A.A.A.; MORAIS, A.R. **Exatidão na classificação de fragmentos de matas em imagem do satélite Cbers-CCD, no município de Lavras, MG**. In: XIII SIMPÓSIO BRASILEIRO DE SENSORIAMENTO REMOTO,

2007, Florianópolis, SC, Brasil. Anais... Florianópolis: INPE/SELPER Brasil, 2007, p. 887-894.

FOODY, G. M. **Status of land cover classification accuracy assessment**. *Remote Sensing of Environment*, v. 80, p. 185–201, 2002.

FORKUOR, G.; CONRAD, C.; THIEL, M.; ULLMANN, T.; ZOUNGRANA, E. **Integration of Optical and Synthetic Aperture Radar Imagery for Improving Crop Mapping in Northwestern Benin, West Africa**. *Remote Sens.* 2014, 6, 6472–6499.

GOETZ S. J. **Crisis in Earth observation**. *Science* 315: 1767, 2007.

GHOSH, S.M.; BEHERA, M.D. **Aboveground biomass estimation using multi-sensor data synergy and machine learning algorithms in a dense tropical forest**. *Appl. Geogr.* 2018, 96, 29–40.

HALL, M. A. **Correlation-based Feature Subset Selection for Machine Learning**. Tese (Doutorado) — University of Waikato, Hamilton, New Zealand, 1998.

HAIR, J. F., ANDERSON, R. E., TATHAM, R. L., BLACK, W. C. **Análise Multivariada de Dados**. Bookman. 2005.

HAN, J.; KAMBER, M.; PEI, J. **Data mining: concepts and techniques: concepts and techniques**. Elsevier, 2011. ISBN 0123814804.

HANSEN M.C., STEHMAN S.V., POTAPOV P.V., LOVELAND T.R., TOWNSHEND J.R.G., *et al.* **Humid tropical forest clearing from 2000 to 2005 quantified by using multitemporal and multiresolution remotely sensed data**. *Proc Nat Acad Sci USA* 105: 9439–9444, 2008.

HANSEN M.C., POTAPOV P.V., MOORE R., HANCHER M., TURUBANOVA S.A., *et al.* **High-resolution global maps of 21st-century forest cover change**. *Science* 342: 850–853, 2013.

HARDOY, J.; MITLIN, D.; SATTERTHWAITTE, D. **Environmental problems in an urbanizing world: Finding solutions in cities in Africa, Asia and Latin America**. London: Routledge. 2013.

HENRIQUES, A. P. M. **Classificação de imagens de ambientes coralinos: uma abordagem empregando uma combinação de classificadores e máquina de vetor de suporte / Antônio de Pádua de Miranda Henriques**. – Natal, RN, 2008. 148 f. : il.

HICKEY, R. **Slope Angle and Slope Length Solutions for GIS**. *Cartography*, v. 29, n. 1, p. 1–8, jun. 2000. Disponível em: <<http://www.tandfonline.com/doi/abs/10.1080/00690805.2000.9714334>>.

HORN, B. K. P. **Hill shading and the reflectance map**. Proceedings of the IEEE, v. 69, n. 1, p. 14–47, 1981. Disponível em: <<http://ieeexplore.ieee.org/document/1456186/>>.

HUANG C., DAVIS L. S. & TOWNSHEND J. R. G. **An assessment of support vector machines for land cover classification**. International Journal of Remote Sensing, 23:4, 725-749, 2002.

HUBER, O.; RIINA, R. (ed.). **Glossário fitoecológico de las Américas. Caracas: Ediciones Tamandúa, Unesco**. 1997. V.1. 500p.

INSTITUTO ESTDUAL DE FLORESTAS (IEF). IEF – INSTITUTO ESTADUAL DE FLORESTAS. Regional Sul: relatório. Varginha, 1996. (Não publicado). Disponível em <http://www.ief.mg.gov.br/index>. Acessado em março de 2020.

JENSEN, J. R. **Sensoriamento remoto do ambiente: uma perspectiva em recursos terrestres**; tradução José Carlos Neves Epiphanyo (coordenador).[et al.]. - São José dos Campos, SP: Parêntese, 2009.

LIAW A., WIENER M. **Classification and regression by random forest**. Forest. 23. 2001.

LOARIE S.R., JOPPA L.N., PIMM S.L. **Satellites miss environmental priorities**. Trends EcolEvol 22: 630–632, 2007.

LU, D.; WENG, Q. **A survey of image classification methods and techniques for improving classification performance**. International Journal of Remote Sensing, v.28, p.823–870, 2007.

MACHADO, R. B. *et al.* **Estimativas de perda da área do Cerrado brasileiro. Brasília: Conservation International do Brasil**, 2004.

MATHER, P. M. **Computer Processing of Remotely Sensed Images**. Wiley, New York, (1999).

MARCUSSI, A. B.; BUENO, C. R. P.; MIQUELONI, D. P.; ARRAES, C. L. **Utilização de índices de vegetação para os sistemas de informação geográfica**. Revista Caminhos da Geografia 11, 41 -53, 2010.

MEDIUM. **Random Forest Simple Explanation**. <https://medium.com/@williamkoehrsen/random-forest-simple-explanation> - (Acessado 14 de setembro de 2020).

MENDEL, J. M. & MCLAREN R. W. **Reinforcement-learning control and pattern recognition systems**. pp. 287–318, 1994.

MENESES, P. R.; ALMEIDA, T. **Introdução ao Processamento de Imagens de Sensoriamento Remoto**. Brasília, 2012.

- MENESES, P.R. & SANO, E.E. **Classificação pixel a pixel de imagens**. In: MENESES, P.R & ALMEIDA T. DE. (Coordenadores), *Introdução ao processamento de imagens de sensoriamento remoto*. Brasília: UNB, CNPQ, p. 191- 208, 2012.
- NOBREGA, R. A. A. **Detecção da malha viária na periferia urbana de São Paulo utilizando imagens de alta resolução espacial e classificação orientada a objetos**. 2007. Tese (Doutorado em Engenharia de Transportes) - Escola Politécnica, Universidade de São Paulo, São Paulo, 2007. doi: 10.11606/T.3.2007.tde-10082007-171032. Acesso em: 2020-11-06.
- NOVO, E. M. L. **Sensoriamento Remoto: princípios e aplicações**. 4.ed. São Paulo: Edgar Blücher, 2010.
- OSUNA, E.E., FREUD, R., *et al.* **Support Vector Machines: Training and Applications**. A.I. Memo No. 1602, C.B.C.L. Paper No. 144. Massachusetts. Institute of Technology and Artificial Intelligence Laboratory, Massachusetts. 1997.
- ORFEO ToolBox. **Open source processing of remote sensing images**. Disponível em: <https://www.orfeo-toolbox.org/>. Acesso em: fevereiro de 2020.
- OTUKEI, J. R. *et al.* **Land cover change assessment using decision trees, support vector machine and maximum likelihood classification algorithms**. International Journal of Applied Earth Observation and Geoinformation 12S (2010) S27–S31.
- PLANO MUNICIPAL DE SANEAMENTO BÁSICO – PMSB. **Produto 3 – Diagnóstico Técnico Participativo**. Indianópolis – MG. DRZ Geotecnologias e Consultoria LTDA. 2014.
- QUINLAN J. R. **Programs for machine learning**. Morgan Kaufmann series in machine learning, San Mateo, Califórnia. 1993.
- QUORA. **How can I use a Support Vector Machine in regression-tasks-SVM**. <https://www.quora.com/>- (Acessado 14 de setembro de 2020).
- RATTER J., RIBEIRO J., BRIDGEWATER S. (1997) **The Brazilian cerrado vegetation and threats to its biodiversity**. Ann Bot 80:223–230.
- RILEY, S.J., DEGLORIA S.D and ELLIOT R. **A terrain ruggedness index that quantifies topographic heterogeneity**, *Intermountain Journal of Sciences*. 5(1-4):23-27, 1999.
- ROSENTHAL, W.D.; BLANCHARD, B.J. **Active microwave responses: An aid in improved crop classification**. *Photogramm. Eng. Remote Sens* 1984, 50, 461–468.

SCHIEWE, J.; TUFTE, L. O potencial de procedimentos baseados em regiões para a avaliação integrada de dados de SIG e sensoriamento remoto. In: **Sensoriamento Remoto e SIG avançados**. Ed. BLASCHKE, T.; KUX, H. 2 ed. São Paulo: Oficina de Textos, 2007. p. 56-65.

SMOLA, A. *et al.* **Advances in large margin classifiers**. London: MIT Press, 2000.

STEVIĆ, D.; D.; B, I.; DOJČINOVIĆ, N.; JOKOVIĆ, J. **Automated identification of land cover type using multispectral satellite images**. Energy and Buildings, v. 115, p. 131-137, 2016.

STRASSBURG B. *et al.* **Moment of truth for the Cerrado hotspot**. *Nature Ecology and Evolution*. 1. 1-3. 10.1038/s41559-017-0099. 2017

TAVARES, P.A.; BELTRÃO, N.E.S.; GUIMARÃES, U.S.; TEODORO, A.C. **Integration of Sentinel-1 and Sentinel-2 for Classification and LULC Mapping in the Urban Area of Belém, Eastern Brazilian Amazon**. *Sensors* 2019, 19, 1140.

TEIXEIRA, M. I. J. G. *et al.* **Florística e fitossociologia de área de Cerrado S.S. no município de Patrocínio Paulista, nordeste do estado de São Paulo**. *Bragantia*, Campinas, v. 63, n. 1, p. 1-11, 2004.

TSO, B.; MATHER, P.M. **Classification methods for remotely sensed data**. London: Taylor and Francis, 2001. 332p.

TURNER W., SPECTOR S., GARDINER N., FLADELAND M., STERLING E., *et al.* **Remote sensing for biodiversity science and conservation**. *Trends Ecol Evol* 18: 306–314, 2003.

USGS/NASA. **United States Geological Survey/National Aeronautics and Space Administration**. 2018. Disponível em: <https://www.usgs.gov/core-science-systems/nli/landsat/landsat-satellite-missions> > acesso em: 12 de setembro de 2020.

VALERIANO, M. M. **Dados Topográficos**. In: FLORENZANO, T. G. (org). *Geomorfologia: conceitos e tecnologias atuais*. São Paulo: Oficina de Textos, 2008. Cap. 3, p.72-103.

XIE Y., SHA Z., YU M. **Remote sensing imagery in vegetation mapping: a review**. *J Plant Ecol* 1: 9–23, 2008.

KULKARNI, A. D. and LOWE, B. **"Random Forest Algorithm for Land Cover Classification"**. Computer Science Faculty Publications and Presentations. Paper 1. 2016.

WILLIAMS D.L., GOWARD S., ARVIDSON T. **Landsat: Yesterday, today, and tomorrow**. *Photogramm Eng Rem Sens* 72: 1171–1178, 2006.

WORLD WIDE FUND FOR NATURE – WWF – Brasil. *Cerrado, Berço das Águas*, 2012. Disponível em:

[https://www.wwf.org.br/natureza\\_brasileira/questoes\\_ambientais/biomas/bioma\\_cerrado/bioma\\_cerrado\\_ameacas/](https://www.wwf.org.br/natureza_brasileira/questoes_ambientais/biomas/bioma_cerrado/bioma_cerrado_ameacas/). Acessado em janeiro de 2020.

YOCKY, D. A. **Multiresolution Wavelet Decomposition Image Merger of Landsat Thematic Mapper and SPOT Panchromatic Data**. Photogrammetric Engineering & Remote Sensing, Vol. 62, No. 9 September 1996, pp. 1067-1074.