

APLICAÇÃO DE MACHINE LEARNING E SENSORIAMENTO REMOTO PARA APRENDIZADO SUPERVISIONADA DA COBERTURA DA TERRA NA SUB-BACIA DO RIO JUQUERI

Ferdinando Yoshio Agapito Urasaki¹

1 INTRODUÇÃO

As técnicas de *Machine Learning* vêm sendo bastante difundidas em diferentes áreas do conhecimento, frente a isso, uma imensa quantidade de dados são geradas diariamente, trazendo consigo uma demanda crescente por métodos eficazes de extração de informação. Neste quesito, o Sensoriamento Remoto se ajusta bem aos algoritmos de *Machine Learning*, possibilitando a automatização de processos.

O Mapeamento do espaço territorial, dando enfase a Cobertura da Terra, é uma importante fonte de informação, base para diversos estudos na área da engenharia, agronomia, socieconomia, meio ambiente, dentre outros. O Sensoriamento Remoto agregado as técnicas de *Machine Learning* permitem obter informações sobre a superfície terrestre através de algoritmos capazes de reconhecer padrões complexos de forma sistemática.

Técnicas de *Machine Learning* vêm sendo cada vez mais utilizadas no Geoprocessamento e no Sensoriamento Remoto, unindo a ciência de dados com a análise espacial, exemplos desta junção podem ser vista no módulo *ENVI Deep Learning*², que trabalha com sensoriamento remoto utilizando algoritmos de aprendizagem profunda; e no *ArcGIS Notebooks*³, que utiliza bibliotecas de geoprocessamento e algoritmos de *Machine Learning* para analise espacial.

Este trabalho tem como objetivo apresentar metodologia de aprendizado supervisionado, aplicando os fundamentos de Sensoriamento Remoto com as técnicas de *Machine Learning*, para construção de modelos preditivos da Cobertura da Terra, visando seu uso em processos semi-automatizados de Classificação utilizando plataforma e softwares livres gratuitos.

O Estudo de Caso foi desenvolvido na área da Sub-Bacia do rio Juqueri situada no Estado de São Paulo, extraindo informações da Imagem de Satélite Landsat TM 5 de órbita/ponto: 219/076 de 24/08/2010 para construção de modelos de aprendizado supervisionado utilizando como referência as Classes do Mapeamento de Cobertura da Terra do Estado de São Paulo (2010) na escala 1:100.000.

Apesar de vasta quantidade de algoritmos de *Machine Learning* para classificação por aprendizado supervisionado, este Estudo de Caso se atreve aos seguintes algoritmos: *Gaussian Naive Bayes*, *Decision Tree*, *Random Forest*, *KNeighbors*, *Logistic Regression*, *Linear Support Vector* e *Multi-Layer Perceptron*.

2 REFERENCIAL TEÓRICO

2.1 Cobertura da Terra

O conhecimento das mudanças na Cobertura da Terra, seja por atividades naturais ou antrópicas, são essenciais para o entendimento da dinâmica ambiental e compreensão das implicações decorrentes da ação antrópica, fundamental para uma adequada gestão ambiental e uso sustentável dos recursos naturais [7].

O termo *Cobertura da Terra* pode ser definido como o estado biofísico da superfície terrestre – diretamente observado ou indiretamente por sensoriamento remoto, identificando-se a assinatura espectral das classes de Cobertura da Terra, não devendo nos confundir com o termo *Uso da Terra* que diz respeito em como são utilizados os elementos biofísicos da superfície e/ou seu propósito [11].

Mudanças na Cobertura da Terra podem gerar impactos ambientais tanto a nível local como global,

¹Eng. Sanitarista e Ambiental, Especialista em Geoprocessamento.

²Extensão do Software de Sensoriamento Remoto *ENVI*, desenvolvido pela *Harris Geospatial*.

³Ferramenta incorporada ao Sistema de Informações Geográficas *ArcGIS*, desenvolvido pela *ESRI*.

a exemplo do desmatamento, onde podemos citar a perda de biodiversidade, degradação do solo, alteração do ciclo hidrológico, mudanças climáticas e aumento da concentração de gases de efeito estufa [2].

A utilização do Sensoriamento Remoto representa assim uma valiosa ferramenta, possibilitando além do mapeamento da Cobertura da Terra, o monitoramento das mudanças na distribuição do seu estado biofísico.

2.2 Sensoriamento Remoto

O Sensoriamento Remoto trata da obtenção de informações sobre a superfície terrestre, pela análise dos dados adquiridos por sensores que não estão em contato com o objeto de investigação [18].

Através do processamento dos dados, capturados pelos sensores orbitais acoplados a satélites artificiais, é possível realizar o levantamento do meio físico e biótico pela análise da assinatura espectral oriundas das bandas obtidas do imageamento, identificando as características dos diferentes alvos e extraíndo informações.

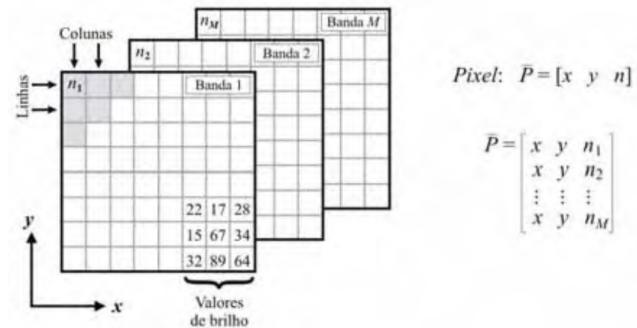
Avanços tecnológicos na área espacial tem promovido o desenvolvimento de sensores orbitais com significantes melhorias nas sua resolução, tanto geométrica quanto espectral, radiométrica e temporal, estando as imagens multiespectrais da superfície terrestre acessíveis em catálogos de imagens, como no caso do Projeto *LANDSAT*⁴ e *CBERS*⁵, disponibilizados de forma gratuita.

Aliado a isso, a velocidade de processamento dos computadores, a capacidade de armazenamento de dados e o desenvolvimento de novas técnicas e algoritmos de extração de informação tiveram avanços significativos, possibilitando que grandes quantidades de dados sejam analisados e processados com rapidez e precisão.

Segundo Richards (1999) [21], uma imagem multiespectral gerada pelos sensores orbitais é constituída por um conjunto de bandas cobrindo uma mesma área, onde a radiação eletromagnética é registrada em diferentes intervalos espectrais. A imagem é arma-

zenada na forma de matriz, onde cada pixel possui coordenadas espaciais (x, y) e respectivo valor de brilho, também denominado de número digital (*digital number – DN*), além de um vetor de atributos cuja dimensão equivale a quantidade de bandas espectrais (apud Prado, 2009 [19]).

Figura 1: Imagem multiespectral obtida por sensores remotos



Fonte: Richard (1999) apud Prado (2009)

2.3 Machine Learning

O Aprendizado de Máquina, mais comumente denominado *Machine Learning* (em inglês), consiste em uma área da Inteligência Artificial que se apresenta como uma ferramenta utilizada na análise de dados, produzindo modelos analíticos de forma automatizada e algoritmos capazes de extrair informações e reconhecer padrões complexos [23].

Assim como a tecnologia, os algoritmos utilizados em *Machine Learning* estão em constante evolução e novos métodos são criados e aperfeiçoados para análise de dados e extração de conhecimento relevantes. Dentre os diversos algoritmos, podemos citar: Gaussian Naive Bayes, Decision Tree, Random Forest, KNeighbors, Logistic Regression, Linear Support Vector e Multi-Layer Perceptron, sendo que a forma de aplicação de *Machine Learning* irá depender do algoritmo adotado e do processo de aprendizado utilizado.

O processo de aprendizado pode ser classificado em Supervisionado, Não-Supervisionado, Semi-Supervisionado e por Reforço, distinguindo-se pela forma como os padrões são detectados. Abordaremos neste Estudo de Caso o Aprendizado Supervisionado,

⁴<https://earthexplorer.usgs.gov/>

⁵<http://www.dgi.inpe.br/catalogo/>

mais especificamente na Classificação da Cobertura da Terra através de Imagem de Satélite Landsat.

2.4 Aprendizado Supervisionada

No Aprendizado Supervisionado algoritmos são utilizados para induzir modelos preditivos, onde é fornecido um conjunto de dados de treinamento em que os rótulos das Classes são previamente conhecidos. O objetivo do Aprendizado Supervisionado é construir um modelo que permita prever as Classes para conjuntos de dados previamente não vistos, ou seja, o conjunto de dados passa por um processo de treinamento que consiste em identificar qual a melhor forma de separar os dados de acordo com os rótulos identificados de modo a ocorrer o menor número de erros possíveis em sua previsão [4].

3 ÁREA DE ESTUDO

A Sub-Bacia do rio Juqueri, com uma área de drenagem de aproximadamente 845 Km^2 , está inserida na Unidade de Gerenciamento de Recursos Hídricos do Alto Tietê (UGRHI-6). Localizada no Estado de São Paulo, a Sub-Bacia engloba os municípios de Nazaré Paulista, Mairiporã, Franco da Rocha, Francisco Morato, Caiearas, São Paulo, Cajamar, Santana de

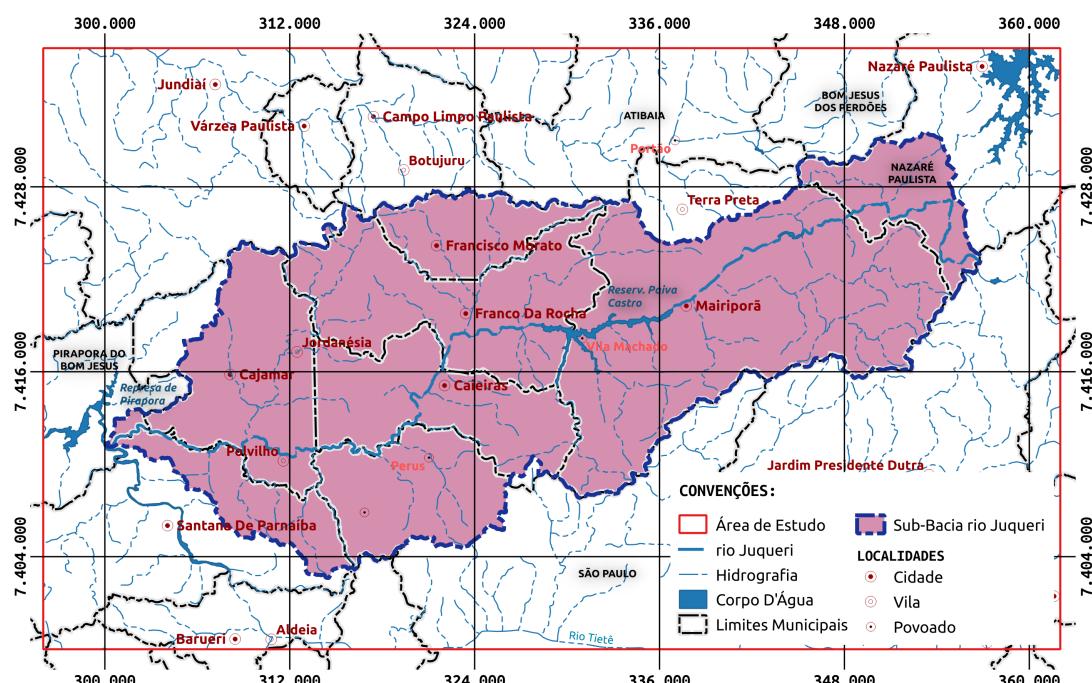
Parnaíba e Pirapora do Bom Jesus.

Com suas nascentes no município de Nazaré Paulista e se estendendo por cerca de 83 Km até sua foz no rio Tietê, o rio Juqueri tem suas águas represadas pela Barragem Paiva Castro, formando o último reservatório do Sistema Cantareira, localizado no município de Mairiporã, contribuindo com aproximadamente 46% do volume de água que abastece a Região Metropolitana de São Paulo (RMSP) [3].

O clima na Sub-Bacia do rio Juqueri se alterna entre estação quente e úmida e outra fria e relativamente seca. No período de julho a outubro as temperaturas variam entre 15.8°C à 19.0°C , com média de 22.4°C no verão. A evaporação média registra valores entre 104.8 mm em agosto e 78.2 mm em julho e a pluviosidade média no ano gira em torno de 1.455 mm , com o mês de janeiro apresentando um valor médio de 238.7 mm (período chuvoso) e agosto com 38.9 mm (período seco) [1].

A Área de Estudo (Figura 2) foi delimitada, envolvendo a Sub-Bacia do rio Juqueri, pelas seguintes coordenadas: canto superior esquerdo (E 296.000 m e N 7.437.000 m) e canto inferior direito (E 362.000 m e N 7.398.000 m). As coordenadas estão referenciadas na projeção Universal Transversa de Mercator (UTM) Datum SIRGAS 2000 Zona 23 Sul.

Figura 2: Sub-Bacia do rio Juqueri



Fonte: produzido pelo autor

4 MATERIAIS

4.1 Classes de Cobertura da Terra

Para o desenvolvimento do Estudo de Caso foi utilizado o Mapeamento de Cobertura da Terra do Estado de São Paulo (2010) na escala 1:100.000. O mapeamento foi realizado pela Coordenadoria de Planejamento Ambiental (CPLA), vinculado à Secretaria de Meio Ambiente do Estado de São Paulo, utilizando classificação baseada em objetos nas imagens Landsat TM 5 com posterior correção visual através de ajuste na classe *Corpos D'água* e nos elementos rodovia e estrada presentes na classe *Área Construída* a partir de bases cartográficas do DAEE⁶ e DER⁷, além da remoção de polígonos menores que 2 hectares (*ha*) classificados como *Área Construída* e *Corpos D'água*, e menores do que 4 *ha* para demais classes. Na validação foram estabelecidos, pela CPLA, um total de 100 pontos amostrais por classe obtendo nível de acertos por classe superior a 90% e acurácia global de 97.14% [8].

As classes abrangidas pelo Mapeamento da Cobertura da Terra e seus respectivos códigos foram:

1. Cobertura Arbórea: Formação vegetal formada predominantemente por elementos arbóreos. matas ciliares, floresta estacional semidecídua, floresta ombrófila densa ou mista, e área de cerrado, mangue e restinga com vegetação de maior porte, além de formações arbóreas homogêneas plantadas como pinus, eucalipto, seringueira e citrus.
2. Cobertura Herbácea-Arbustiva: Formação herbácea e/ou arbustiva, solo coberto por vegetação de gramíneas ou leguminosas, áreas de pasto melhoradas ou cultivadas destinadas a pastoreio, culturas temporárias, semi-perenes e perenes, terras cultivadas tanto como zonas agrícolas heterogêneas ou homogêneas e áreas remanescentes de cerrado e restinga.
3. Solo Exposto: Área de intervenção antrópica que

foram terraplenadas ou aradas, incluindo áreas em transição de uso ou em fase intermediária de uso ou em processo erosivo com exposição do solo.

4. Área Construída: Áreas de uso intensivo, com edificações e sistema viário, com predomínio de superfícies artificiais não-agrícolas.
5. Áreas Úmidas: Áreas com lençol freático na superfície, estabelecendo geralmente vegetação aquática. Inclui áreas de brejos, pântanos e extensas áreas junto às várzeas ou áreas de mineração inundáveis.
6. Corpos D'água: Águas interiores como cursos d'água, rios, riachos, canais, lagos naturais e reservatórios artificiais.
7. Sombra e Nuvem: Inclui áreas cobertas com nuvens e sombras.

4.2 Imagem de Satélite Landsat

O programa *LANDSAT* surgiu no final da década de 60, como parte do Programa de Levantamento de Recursos Terrestres da *NASA*⁸ em conjunto com outras agências federais dos EUA, fornecendo dados multiespectrais de média resolução da superfície da Terra em uma base global, acessível livremente ao público [9].

Na aquisição da cena foi selecionada imagem proveniente do sensor TM (*Thematic Mapper*) do Satélite Landsat 5 correspondente a órbita/ponto: 219/076 de 24/08/2010, disponibilizada gratuitamente pela *USGS*⁹ *EROS*¹⁰.

Adotou-se como critérios de seleção as imagens que possuíssem baixa cobertura de nuvens (inferior a 10%), imageadas no período seco, no ano de 2010.

O ano de 2010 foi definido com intuito de compatibilizar a assinatura espectral da imagem de satélite com as Classes do Mapeamento de Cobertura da Terra, de mesmo ano, utilizado como referência.

⁶Departamento de Águas e Energia Elétrica

⁷Departamento de Estradas e Rodagem

⁸National Aeronautics and Space Administration

⁹United States Geological Survey

¹⁰Earth Resources Observation and Science (<http://earthexplorer.usgs.gov/>)

O Landsat TM 5, lançado em março de 1984, foi o que mais tempo atuou no programa *LANDSAT*, sendo desativado em novembro de 2011 após 29 anos em operação. As cenas tinham faixa imageada de 185 Km, e possuíam resolução temporal de 16 dias (período de revisita) e radiométrica de 8 bits (256 níveis de cinza) [24].

Tabela 1: Bandas do Sensor Landsat TM 5

BANDAS	COMPRIMENTO DE ONDA (um)	RESOLUÇÃO ESPACIAL (m)
Banda 1 - Azul	0.45 - 0.52	30
Banda 2 - Verde	0.52 - 0.60	30
Banda 3 - Vermelho	0.63 - 0.69	30
Banda 4 - Infravermelho Próximo (NIR)	0.76 - 0.90	30
Banda 5 - Infravermelho Médio(SWIR 1)	1.55 - 1.75	30
Banda 6 - Infravermelho Termal (TIR)	10.40 - 12.50	120
Banda 7 - Infravermelho Médio (SWIR 2)	2.08 - 2.35	30

Fonte: USGS (2020)

4.3 Softwares

Os trabalhos foram desenvolvidos utilizando a plataforma do Sistema Operacional Linux 64 bits (Kernel 5.4), distribuição Ubuntu v.20.04. As ferramentas adotadas fazem parte do rol dos softwares livres gratuitos, destacando-se: QGIS LTR v3.10, SCP – Semi-Automatic Classification Plugin v6.4.6 (complemento do QGIS), Spyder v4.1.3 e as bibliotecas Python: numpy, pandas, pyrsgis, imblearn, sklearn, hyperopt e pickle.

4.4 Hardware

A configuração do hardware utilizada consiste em Desktop equipado com processador de 6 núcleos rodando a 3.5GHz (AMD FX-6300), 14Mb de cache, 24Gb de memória ram (DDR3 1.866Mhz), placa gráfica de 2Gb GDDR5 a 128bits (NVidea GTX 750Ti) e armazenamento interno SSD de 512Gb.

5 METODOLOGIA

A Metodologia utilizada neste Estudo de Caso foi adaptado da *CRISP-DM*¹¹, esta metodologia é composta de seis etapas que direcionam a descoberta do conhecimento para tomada de decisão: *Entendi-*

mento do Problema (determinação dos objetivos e necessidades); *Compreensão dos Dados* (identificação das informações relevantes); *Preparação dos Dados* (envolve atividades de integração de dados, obtendo um conjunto de dados que reflete as necessidades dos algoritmos de aprendizagem); *Modelagem* (consiste na escolha dos algoritmos e seus parâmetros para construção do modelo); *Avaliação* (refere-se a verificação dos objetivos e resultados obtidos) e *Implementação* (envolve colocar o modelo em produção para que possa ser utilizado). [25].

As etapas de *Entendimento do Problema* e *Compreensão dos Dados* já foram abordados nas seções anteriores, e seguindo um encadeamento lógico, trataremos da fase de *Preparação dos Dados*.

5.1 Preparação dos Dados

A preparação dos dados pode envolver diversas operações e está intimamente ligada a capacidade de se identificar os problemas presentes e escolher os métodos apropriados para solucionar cada um deles.

Neste Estudo de Caso, a preparação dos dados para Classificação da Cobertura da Terra através do Aprendizado Supervisionado consiste nas atividades de pré-processamento (correção atmosférica, conversão em valores de reflectância, empilhamento de bandas, transformação da projeção cartográfica, recorte e rasterização), limpeza de dados (correção de registros incompletos, incorretos e inconsistentes), e tratamento de dados desbalanceados, de forma a se obter um conjunto de dados propício a etapa de modelagem.

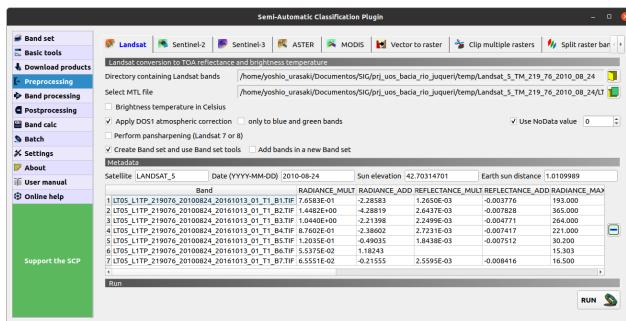
O produto final desta etapa consiste em um conjunto de dados (*Dataset*), limitados espacialmente pela área de estudo, sendo formado pela integração dos dados provenientes da imagem de satélite (matriz de n bandas com dados de reflectância com correção atmosférica) e do mapeamento da cobertura da terra (matriz com dados da classe de cobertura da terra codificado na forma de variável numérica não ordinal), com remoção dos registros referentes a classe *Sombra* e *Nuvem* e ruídos presentes na região de fronteira entre classes.

¹¹Cross Industry Standard Process for Data Mining

5.1.1 Pré-processamento da Imagem de Satélite

a) **Correção Atmosférica:** Tem o objetivo de reduzir os efeitos provenientes da interferência atmosférica na imagem de satélite. Das técnicas que visam reduzir os efeitos de dispersão atmosférica optou-se por utilizar o *DOS (Dark Object Subtraction)*. Esta técnica de Subtração por Pixel Escuro é amplamente utilizada pela simplicidade na determinação de seus parâmetros. O valor medido em pontos como lagos, água limpa e sombra, onde a radiância teoricamente é nula, é atribuíndo à interferência atmosférica, sendo subtraído de toda a cena [13].

Figura 3: Plugin SCP no QGIS – DOS and TOA Reflectance Conversion



Fonte: produzido pelo autor

b) **Conversão de Números Digitais em Valores de Reflectância:** Os valores físicos geralmente convertidos dos números digitais são a reflectância (*adimensional*) e a radiância ($mW \cdot cm^{-2} \cdot sr^{-1}$), possibilitando a comparação de determinada feição e a identificação de mudanças na sua resposta espectral ao longo do tempo.

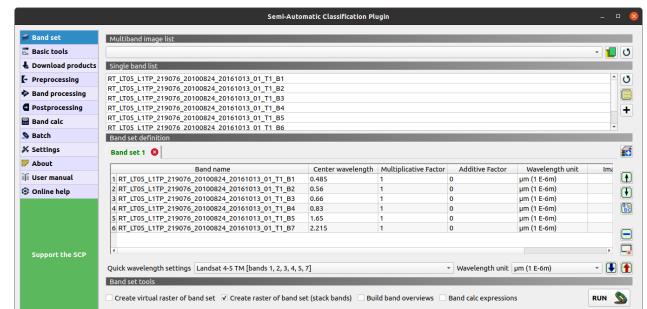
Os Números Digitais (valores digitais brutos) são utilizados na classificação de imagens para diversos fins porém estes valores não representam quantitativamente variáveis físicas reais como a reflectância ou radiância e portanto não podem ser comparados entre sensores distintos ou imagens com datas de imageamento diferente devido a mudança na calibração dos sensores e fatores como o ângulos de iluminação solar em que cada imagem é obtida [17].

Deste modo foi realizada a conversão dos valores

dos pixels da Imagem de Satélite em Reflectância do Topo da Atmosfera (*TOA Reflectance*).

c) **Empilhamento de Bandas (Stack Bands):** Consiste na seleção das bandas que serão utilizadas no aprendizado supervisionado. O processo faz o empilhamento das bandas, onde inicialmente cada arquivo *GeoTIFF*¹² representa uma única banda, e retorna um *GeoTIFF* com todas as bandas selecionadas. As bandas 1, 2, 3, 4, 5 e 7 foram empilhadas para formar um único arquivo Raster.

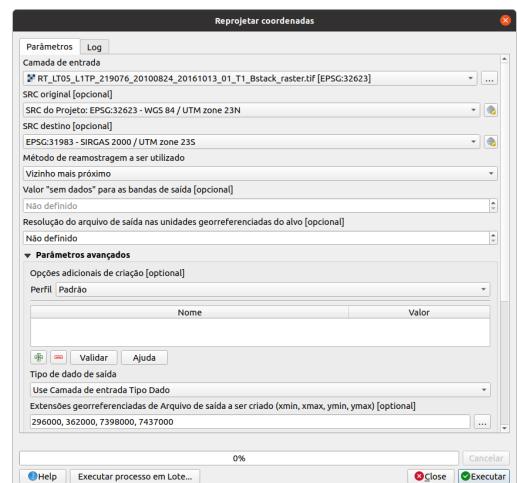
Figura 4: Tela do plugin SCP no QGIS – Stack Bands



Fonte: produzido pelo autor

d) **Reprojecção do Sistema de Coordenadas:** Redefine a Projeção Cartográfica da Imagem de Satélite, de WGS 84 UTM Zona 23N para SIRGAS 2000 UTM Zona 23S, necessário para se adequar a legislação¹³, que determina o SIRGAS 2000 como referencial geodésico a ser adotado na produção cartográfica brasileira.

Figura 5: QGIS – Ferramenta gdal:warpereproject e Recorte



Fonte: produzido pelo autor

¹²Geographic Tagged Image File Format

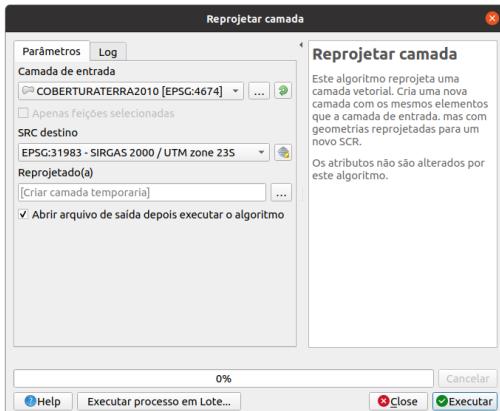
¹³Resolução do Presidente do IBGE nº 1, de 25/02/2005

e) **Recorte da Imagem de Satélite:** Esta última etapa do Pré-processamento da Imagem de Satélite é responsável por recortar a área de interesse baseado nos limites da Área de Estudo. A saída consiste em um arquivo matricial no formato *GeoTIFF* contendo as bandas 1, 2, 3, 4, 5 e 7, em que cada pixel da matriz tem seu valor respectivo de reflectância do topo da atmosfera.

5.1.2 Pré-processamento da Cobertura da Terra

a) **Reprojeção do Sistema de Coordenadas:** Redefine a Projeção Cartográfica do Mapeamento de Cobertura da Terra, de GCS SIRGAS 2000 para SIRGAS 2000 UTM Zona 23S, necessário para compatibilizar o mapeamento com o sistema de coordenadas cartesianas utilizada na imagem de satélite.

Figura 6: QGIS – Ferramenta native:reprojectlayer



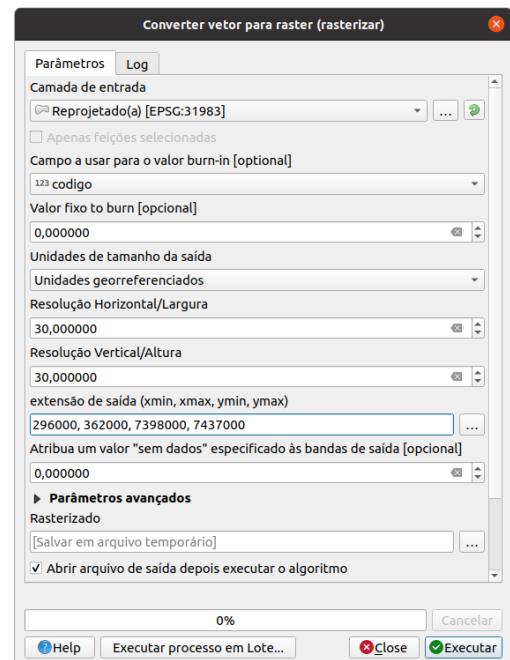
Fonte: produzido pelo autor

b) **Recorte da Cobertura da Terra:** Responsável por selecionar as informações contidas na região de interesse baseado nos limites da Área de Estudo.

c) **Conversão Vetor para Raster:** Denominado de Rasterização, este processo consiste na tarefa de converter geometrias (vetor) em pixels (raster). A saída será um arquivo matricial no formato *GeoTIFF*, em que cada pixel da matriz tem seu valor respectivo ao constante no campo especificado na tabela de atributos da geometria. Para a rasterização do Mapeamento de Cobertura da Terra utilizou-se o campo *codigo* para

os valores dos pixels, este campo contém uma representação numérica, variando de 1 a 7, da variável categórica nominal.

Figura 7: QGIS – Ferramenta gdal:rasterize e Recorte



Fonte: produzido pelo autor

5.1.3 Limpeza de Dados

Dados podem conter irregularidades ou estarem corrompidos, comprometendo a qualidade do conjunto de dados e afetando o desempenho de modelos de aprendizado. Sendo assim, é fundamental analisá-los a fim de identificar problemas e decidir a melhor abordagem para limpeza de dados. Dentre os problemas que podem ser sanados com a limpeza de dados tem-se: registros incompletos (ausência de valores), incorretos (presença de ruído) e inconsistentes (valores conflitantes ou discrepantes) [4].

Identificou-se que apesar do Mapeamento de Cobertura da Terra possuir a Classe *Sombra e Nuvem*, código 7, ela não representa o tipo de cobertura da terra no local demarcado na imagem de satélite (registro incorreto), sendo possível de ser removido do conjunto de dados, deste modo, todos os registros associados a esta classe foram removidos.

Figura 8: Representação da Remoção da Classe Sombra e Nuvem

Imagen de Satélite						Cobertura da Terra
B1	B2	B3	B4	B5	B7	codigo
0.355	0.188	0.233	0.898	0.986	0.264	1
0.033	0.607	0.853	0.229	0.639	0.576	2
0.564	0.418	0.168	0.243	0.302	0.856	5
0.579	0.616	0.363	0.849	0.930	0.315	7
0.307	0.292	0.396	0.199	0.858	0.638	6
0.511	0.922	0.734	0.886	0.933	0.117	2
0.342	0.168	0.381	0.219	0.078	0.552	7
0.057	0.109	0.231	0.775	0.442	0.237	4
0.210	0.116	0.304	0.021	0.192	0.490	1
0.865	0.072	0.034	0.902	0.809	0.645	2

Fonte: produzido pelo autor

5.1.4 Tratamento de Dados Desbalanceados

Define-se como Dados Desbalanceados sempre que a contabilização das observações pertencentes a uma das classes exceda as demais, causando uma grande desproporção entre as distribuições das classes a ponto de exibir uma tendência na classificação para a classe majoritária em detrimento da capacidade do algoritmo em discriminar a classe minoritária. A não verificação desse desequilíbrio pode levar a uma redução significativa no desempenho do classificador em identificar as classes minoritárias, problema que se agrava nos casos em que a previsão da classe minoritária é extremamente relevante [4].

Segundo Machado (2007) [12], para mitigar o impacto negativo do problema de desbalanceamento, inúmeras abordagens foram propostas na literatura, inclusive o uso conjunto de duas ou mais soluções. De forma didática pode-se separar esses métodos em:

- a) **Método de Amostragem de Dados:** Altera a distribuição dos dados com objetivo de aumentar a acurácia de seus modelos.
 - (i) Eliminando os casos de classe majoritária através da aplicação de técnicas de *undersampling* (havendo risco de eliminar dados potencialmente úteis);
 - (ii) Replicando casos da classe minoritária, *oversampling*, agregando cópias exatas das classes minoritárias, porém correndo-se o risco de aumentar a probabilidade de ocorrência de *overfitting*;
 - (iii) Utilizando heurística para eliminar casos de classe majoritária ou replicação de casos

da classe minoritária. ex: CNN (*Condensed Nearest Neighbor Rule*), OSS (*One-Sided Selection*), SMOTE (*Synthetic Minority Oversampling Technique*).

- b) **Método de Limpeza de Dados:** Tenta criar um conjunto de treinamento consistente, removendo os ruídos, valores redundantes e próximos da borda de decisão sem se ater a balancear o conjunto de dados. ex: Tomek Links, ENN (*Edited Nearest Neighbor Rule*), NCL (*Neighborhood Cleaning Rule*), RENN (*Repeated Edited Nearest Neighbor Rule*).

Verificou-se, pela Tabela 2, que o conjunto de dados do Mapeamento de Cobertura da Terra é formado por classes desbalanceadas, sendo relevante seu tratamento com objetivo de obter melhores resultados. Denota-se ainda que as classes do Mapeamento de Cobertura da Terra tem igual importância, essa premissa é necessária para identificar qual solução utilizar para o tratamento dos dados desbalanceados.

Tabela 2: Distribuição das Classes de Cobertura da Terra

CÓDIGO	CLASSE	REGISTROS	%
1	Cobertura Arbórea	1.437.953	50.28
2	Cobertura Herbácea-Arbustiva	489.591	17.12
3	Solo Exposto	144.875	5.07
4	Área Construída	743.320	25.99
5	Áreas Úmidas	4.931	0.17
6	Corpos D'água	36.636	1.28

Assim, optou-se pelo método RENN (*Repeated Edited Nearest Neighbor*), técnica projetada para funcionar como um filtro de ruídos, eliminando dados na região de fronteira (entre classes). O algoritmo é aplicado de forma sucessiva até que não seja possível remover mais nenhum ruído [12].

Tabela 3: Distribuição das Classes de Cobertura da Terra após aplicação do RENN

CÓDIGO	CLASSE	REGISTROS	%
1	Cobertura Arbórea	968.907	63.73
2	Cobertura Herbácea-Arbustiva	50.620	3.33
3	Solo Exposto	9.430	0.62
4	Área Construída	473.990	31.17
5	Áreas Úmidas	4.931	0.32
6	Corpos D'água	12.540	0.82

5.2 Modelagem

A Modelagem consiste na construção do modelo, e envolve a seleção dos algoritmos e a determinação de seus hiperparâmetros.

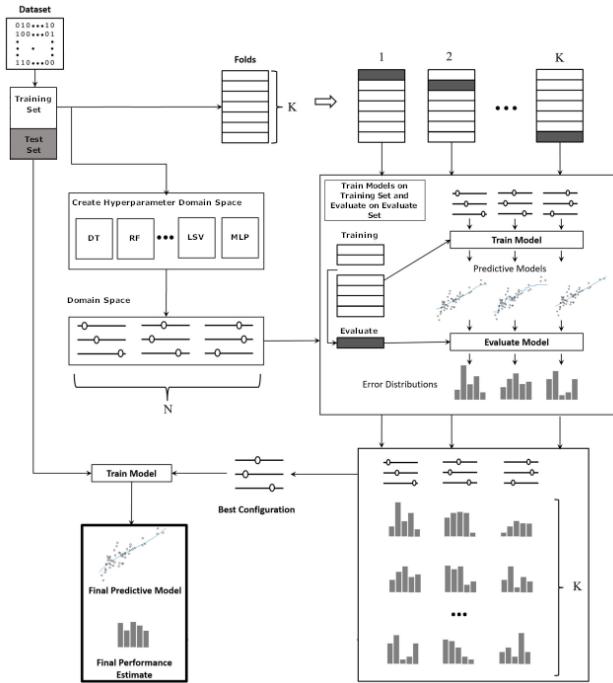
Nesta etapa, diversas técnicas foram utilizadas para se construir os modelos de predição utilizando algoritmos de *Machine Learning*.

Iniciou-se com a divisão do conjunto de dados para que parte fosse utilizada no treinamento do algoritmo e parte para sua validação.

Em seguida foi implantado método de otimização para busca dos hiperparâmetros que se ajustem melhor a cada algoritmo, utilizando o conjunto de dados de treinamento.

Por fim, a construção dos modelos de predição, utilizando os hiperparâmetros otimizados, e sua avaliação.

Figura 9: Esquema da Aplicação de *Machine Learning* para Análise Preditiva



Fonte: adaptado de Borboudakis et.al (2017) [5]

5.2.1 Partição do Conjunto de Dados

Um dos problemas que podem ocorrer ao se otimizar um algoritmo de *Machine Learning* é gerar um modelo que se ajuste muito bem ao conjunto de

dados (*dataset*) mas que se mostre ineficaz para prever novos resultados, este sobre-ajuste é denominado *overfitting*.

Um modo de contornar este problema, e que foi utilizado neste Estudo de Caso, é dividir os dados em dois conjuntos diferentes: um para ajuste do modelo (*training set*) e outro para teste (*test set*), considerando-se como melhor prática que o conjunto utilizado para auxiliar na construção de um modelo não deva ter registros em comum ao conjunto que irá avaliar o seu potencial. Aplicou-se a técnica de partição do conjunto de dados: *Validação Holdout*, sendo definido a proporção de 70% para o conjunto de treinamento e 30% para teste, de forma estratificada (distribuição igual das diferentes classes) [15].

5.2.2 Otimização dos Hiperparâmetros

Segundo Santos (2018) [22], os algoritmos de *Machine Learning* possuem um ou mais parâmetros de entrada que descrevem a complexidade do modelo, estes parâmetros são denominados *hiperparâmetros* e são definidos antes de sua aplicação, não havendo uma fórmula analítica que auxilie na determinação dos valores para se extrair o melhor desempenho do modelo, tendo em vista a variabilidade de configurações possíveis.

Uma solução que permite se ter a melhor configuração dos hiperparâmetros para a construção do modelo consiste na introdução de uma função que auxilie na otimização das buscas dos hiperparâmetros de modo a obter alta performance e boas métricas no treinamento.

Adotou-se para este Estudo de Caso o pacote *Hyperopt* da biblioteca *Sklearn* e o algoritmo de otimização *Tree Parzen Estimator (TPE)*, que utiliza *otimização bayesiana*¹⁴, para reduzir o tempo de processamento, percorrendo de forma mais eficaz o domínio de busca dos hiperparâmetros.

Após definição do domínio de busca dos hiperparâmetros (*hyperparameter domain space*), para cada algoritmo analisado, foi aplicado a técnica de *Validação Cruzada K-Fold* utilizando como métrica o *f1-score* para comparar as configurações dos hiperparâ-

¹⁴Método que utiliza probabilidade condicional gerando soluções otimizadas a cada iteração

metros. O conjunto de dados de treinamento (*training set*) foi dividido randomicamente em 5 grupos (*folds*) do qual um grupo é utilizado para avaliar o modelo (*evaluate set*) e os demais ($k-1$) para treinamento, sendo o processo repetido a cada grupo até que todos os grupos fossem utilizados na avaliação, retornando a média dos *f1-score* obtidos.

O processo se repetiu por 100 interações ou até extrapolar o tempo de execução, definido para 12 horas, de modo que a cada interação as configurações dos hiperparâmetros foram sendo ajustadas. As melhores configurações dos hiperparâmetros obtidos são apresentadas na Tabela 4.

Tabela 4: Resultados obtidos na Otimização dos Hiperparâmetros

ALGORITMO	HIPERPARÂMETROS	DOMÍNIO DE BUSCA	MELHOR CONFIGURAÇÃO
Decision Tree	var smoothing	np.logspace(0, -9, num=10)	1e-2
	criterion	gini, entropy	gini
	splitter	best, random	best
	min samples split	range(2, 10)	2
	min samples leaf	range(1, 10)	1
	min weight fraction leaf	np.logspace(-2, -7, num=6)	1e-7
	max features	sqrt, log2	log2
	min impurity decrease	np.logspace(-2, -7, num=6)	1e-7
	class weight	none, balanced	balanced
Random Forest	n estimators	range(10, 100, 5)	75
	criterion	gini, entropy	gini
	max depth	range(0, 30)	23
	min samples split	range(2, 10)	5
	min samples leaf	range(1, 10)	1
	min weight fraction leaf	np.logspace(-2, -7, num=6)	1e-7
	max features	sqrt, log2	log2
	min impurity decrease	np.logspace(-2, -7, num=6)	1e-7
	bootstrap	true, false	true
KNeighbors	class weight	none, balanced	balanced
	n neighbors	range(1, 10)	2
	weights	uniform, distance	distance
	algorithm	auto, ball tree, kd tree, brute	auto
	leaf size	range(1, 30)	24
Logistic Regression	p	range(1, 7)	6
	penalty	l1, l2	l2
	tol	np.logspace(-2, -7, num=6)	1e-4
	C	hp.uniform('C', 0.01, 100)	71.66
	class weight	none, balanced	balanced
	solver	newton-cg, lbfgs, sag	lbfgs
	max iter	range(100, 500, 20)	220
Linear Support Vector	multi class	ovr	ovr
	penalty	l1, l2	l2
	loss	hinge, squared hinge	squared hinge
	tol	np.logspace(-2, -7, num=6)	1e-4
	C	hp.uniform('C', 0.01, 100)	5.33
Multi-Layer Perceptron	multi class	ovr	ovr
	class weight	none, balanced	balanced
	hidden layer sizes	(10, 10, 10,)	(10, 10, 10,)
	activation	tanh, relu, logistic	tanh
	solver	sgd, adam, lbfgs	adam
Multi-Layer Perceptron	alpha	np.logspace(-2, -7, num=6)	1e-5
	tol	np.logspace(-2, -7, num=6)	1e-4
	learning rate	constant, adaptive	adaptive

5.2.3 Construção dos Modelos de Predição

De posse das configurações dos hiperparâmetros otimizados, encontradas para cada algoritmo, construiu-se os modelos de predição, utilizando-se o conjunto de dados de treinamento.

Salienta-se que ao se utilizar dados não balanceados torna-se importante comparar a qualidade de predição dos modelos sobre métricas estatísticas diferentes. Tais análises buscam medir o desempenho do modelo, avaliando sua capacidade de reproduzir o valor observado.

De acordo com Kubat et al. (1998) [10], é desejável utilizar uma medida de desempenho diferente da *acurácia*, quando se trabalha com dados não平衡ados visto que alguns classificadores são induzidos a serem mais precisos para a classe majoritária (apud Carrijo, 2004 [6]).

Tabela 5: Valores de Referência do Índice *Kappa*

Índice Kappa	Desempenho
≤ 0	Péssimo
$0.00 < K \leq 0.20$	Ruim
$0.20 < K \leq 0.40$	Razoável
$0.40 < K \leq 0.60$	Bom
$0.60 < K \leq 0.80$	Muito Bom
$0.80 < K \leq 1.00$	Excelente

Fonte: Ribeiro (2017) [20]

Neste sentido, o índice *Kappa* é uma métrica considerada robusta para avaliação de dados desbalanceados, por levar em consideração a possibilidade de a concordância ocorrer por acaso, sendo bastante utilizada na área de detecção remota [14].

A Tabela 6 apresenta as métricas *f1-score*, *kappa*, *roc auc*, *accuracy*, *precision* e *recall* resultantes do treinamento dos algoritmos.

Tabela 6: Métricas dos Modelos para o Conjunto de Dados de Treinamento

MODELOS	f1 score	kappa	roc auc	accuracy	precision	recall
Gaussian Naive Bayes	0.9730	0.9395	0.9805	0.9697	0.9775	0.9697
Decision Tree	0.9925	0.9810	0.9952	0.9905	0.9959	0.9905
Random Forest	0.9975	0.9947	0.9986	0.9974	0.9976	0.9974
KNeighbors	0.9999	0.9998	0.9999	0.9999	0.9999	0.9999
Logistic Regression	0.9793	0.9574	0.9863	0.9787	0.9811	0.9787
Linear Support Vector	0.9812	0.9670	0.9875	0.9837	0.9809	0.9837
Multi-Layer Perceptron	0.9930	0.9868	0.9941	0.9935	0.9929	0.9927

Das métricas apresentadas podemos concluir que todos os modelos tiveram excelentes resultados, se ajustando bem ao conjunto de dados de treinamento, *KNeighbors* obteve destaque com índice *Kappa* de 0.9998 enquanto *Gaussian Naive Bayes* obteve o menor índice dos algoritmos analisados, com *Kappa* de 0.9395, mesmo assim considerado de desempenho excelente.

5.3 Avaliação

Nesta etapa existe a preocupação em medir o grau de acerto da predição, analisando o quanto os valores preditos se afastam do conjunto de dados de teste, de modo a indicar qual ou quais modelos de predição tiveram os melhores resultados.

Para verificar se os modelos treinados são capazes de classificar dados não conhecidos, foi aplicado aos modelos o conjunto de dados de teste.

Tabela 7: Métricas dos Modelos para o Conjunto de Dados de Teste

MODELOS	f1 score	kappa	roc auc	accuracy	precision	recall
Gaussian Naive Bayes	0.9728	0.9393	0.9805	0.9696	0.9773	0.9696
Decision Tree	0.9947	0.9872	0.9968	0.9936	0.9966	0.9936
Random Forest	0.9978	0.9955	0.9988	0.9978	0.9980	0.9978
KNeighbors	0.9999	0.9998	0.9999	0.9999	0.9999	0.9999
Logistic Regression	0.9788	0.9572	0.9861	0.9787	0.9805	0.9787
Linear Support Vector	0.9812	0.9670	0.9875	0.9837	0.9809	0.9837
Multi-Layer Perceptron	0.9920	0.9851	0.9935	0.9927	0.9920	0.9927

Analizando os resultados podemos concluir que todos modelos tiveram desempenho excelente. *KNeighbors* obteve 0.9998 de índice *Kappa* (mesmo resultado obtido com os dados de treinamento) e *Gaussian Naive Bayes* obteve o menor índice *Kappa* com 0.9393.

Valores tão expressivos levantam o questionamento se a remoção no *Dataset* dos ruídos presentes na região de fronteira (aumentando a margem de separação entre classes diferentes), pelo método *RENN* (*Repeated Edited Nearest Neighbor*), são os responsáveis por esses relevantes resultados.

Como avaliação final, para Classificação da Cobertura da Terra, foi aplicado aos modelos de predição (agora já treinados) o *dataset* obtido na etapa de *Pré-Processamento*, sendo apresentado na Tabela 8 as métricas obtidas.

Tabela 8: Métricas do Modelos para o Conjunto de Dados obtidos do Pré-Processamento

MODELOS	f1 score	kappa	roc auc	accuracy	precision	recall
Gaussian Naive Bayes	0.7303	0.5909	0.8029	0.7425	0.7240	0.7425
Decision Tree	0.8022	0.6894	0.8675	0.7833	0.8489	0.7833
Random Forest	0.8223	0.7216	0.8784	0.8109	0.8472	0.8109
KNeighbors	0.9375	0.9035	0.9448	0.9389	0.9403	0.9389
Logistic Regression	0.7527	0.6053	0.8236	0.7378	0.7694	0.7378
Linear Support Vector	0.7454	0.6250	0.8208	0.7634	0.7384	0.7634
Multi-Layer Perceptron	0.7655	0.6500	0.8288	0.7808	0.7608	0.7808

Apesar da redução já esperada nas métricas, *KNeighbors* se manteve com desempenho Excelente, com índice *Kappa* de 0.9035, valor superior ao obtido pelo segundo colocado, *Random Forest* com 0.7216, apresentando um desempenho um pouco abaixo.

O resultado da aplicação do Modelo de Aprendizado Supervisionado utilizando o algoritmo *KNeighbors* para Classificação da Cobertura da Terra na Sub-Bacia do rio Juqueri (Figura 10), pode ser comparado com o Mapeamento da Cobertura da Terra na Sub-Bacia do rio Juqueri (Figura 11).

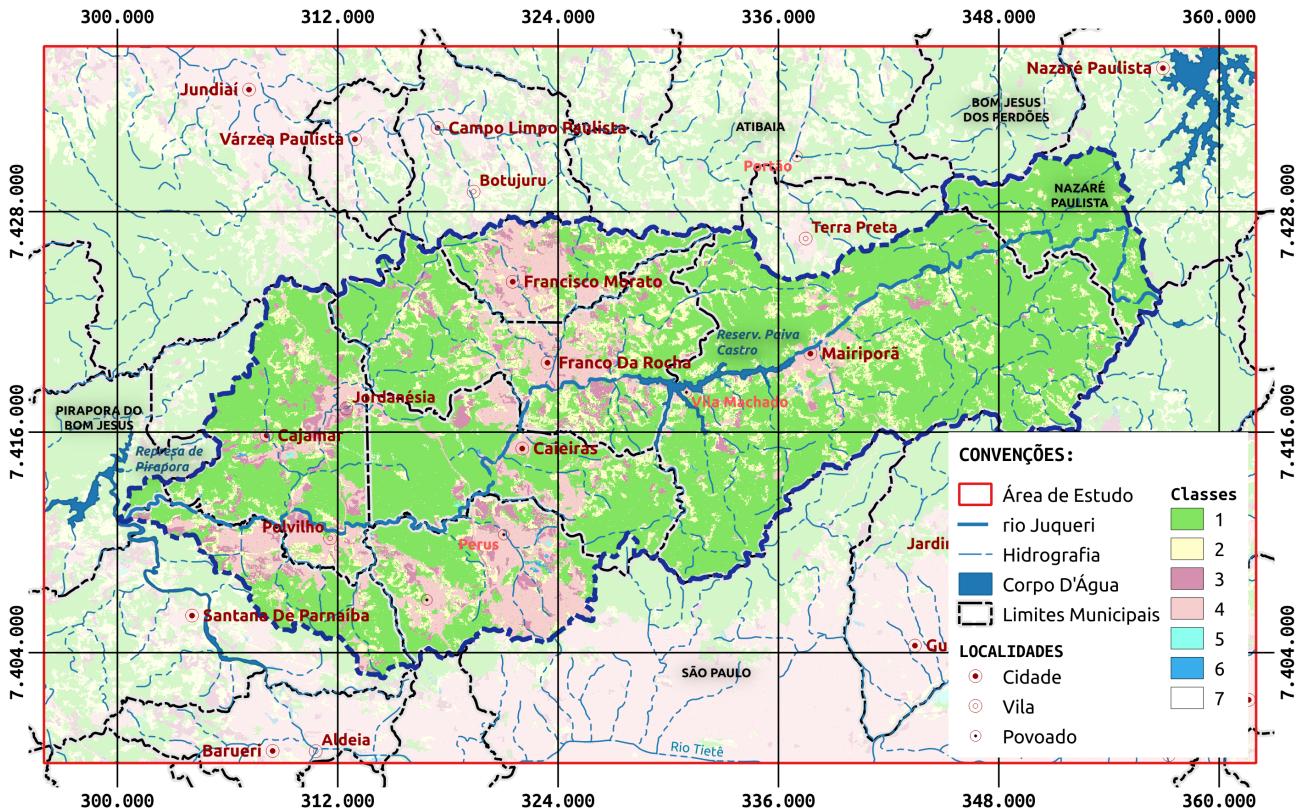
Na Tabela 9, é apresentada a Matriz de Confusão, em porcentagem, para a Classificação da Cobertura da Terra na Sub-Bacia do rio Juqueri.

De acordo com a matriz de confusão da classificação com o modelo *KNeighbors*, a classe 1 (Cobertura Arbórea) foi a que apresentou maior confusão entre as classes utilizadas, sendo que 91.56% dos pixels foram corretamente classificados, obtendo uma confusão na ordem de 6.26% para a classe 2 (Cobertura Herbácea-Arbustiva). A classe 4 (Área Construída) foi a que obteve o maior percentual de acerto (99.72%).

Tabela 9: Matriz de Confusão obtido do Modelo *KNeighbors*

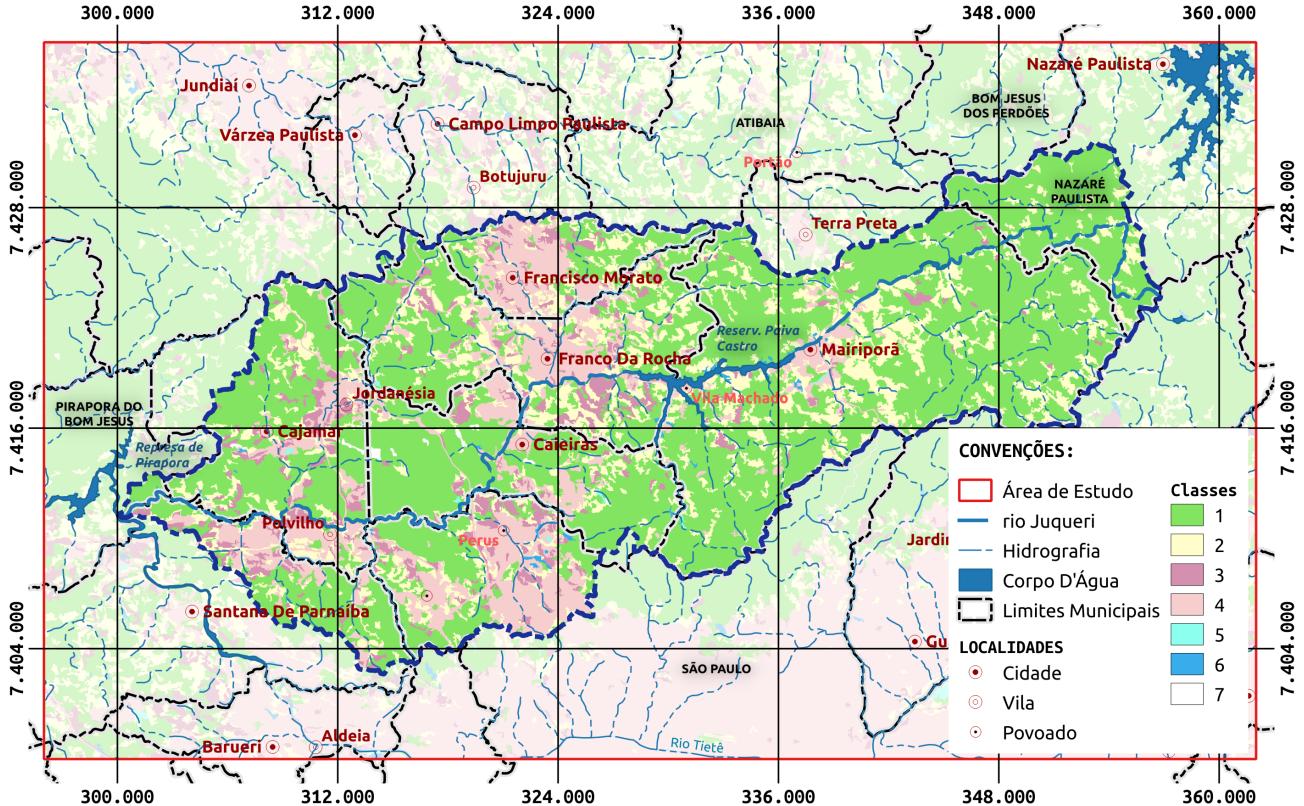
Classes	1	2	3	4	5	6	7
1	91.56	2.17	0.30	0.05	0.57	0.75	1.34
2	6.26	92.18	0.31	0.07	0.09	0.23	0.10
3	0.73	2.09	94.22	0.07	0.03	0.05	0.00
4	0.89	3.24	4.97	99.72	0.00	0.05	0.00
5	0.08	0.06	0.03	0.01	97.73	0.06	0.00
6	0.37	0.25	0.16	0.08	1.57	98.81	0.00
7	0.10	0.01	0.01	0.00	0.00	0.05	98.56
\sum	100.00	100.00	100.00	100.00	100.00	100.00	100.00

Figura 10: Aplicação do Modelo de Predição *KNeighbors* da Cobertura da Terra na Sub-Bacia do rio Juqueri



Fonte: produzido pelo autor

Figura 11: Mapeamento da Cobertura da Terra na Sub-Bacia do rio Juqueri (2010)



Fonte: adaptado de CPLA, 2013 [8]

5.4 Implementação

Segundo Zimmermann et al. (2009) [26], um modelo de predição é transferível para outros projetos quando os valores da precisão, sensibilidade e acurácia são maiores que 0.75 (apud Paula, 2016 [16]).

Assim sendo, considerando as métricas obtidas, os modelos de predição *KNeighbors*, *Random Forest*, *Decision Tree* e *Multi-Layer Perceptron* poderiam ser utilizados para Classificação da Cobertura da Terra em outras áreas de estudo, com destaque ao modelo *KNeighbors* que obteve o melhor desempenho dentre os modelos testados.

Mesmo com bons resultados, os modelos de predição não foram colocados em produção para serem testados em outras áreas, apesar de tecnicamente promissor, por estar fora do escopo deste estudo de caso.

6 CONCLUSÃO

O objetivo principal, alcançado neste trabalho, esteve concentrado em apresentar uma metodologia

de aprendizado supervisionado capaz de unir as especificidades das imagens de satélite e do próprio sensoriamento remoto, e o uso de algoritmos de *machine learning* com seus hiperparâmetros otimizados para construção de modelos preditivos para classificação da cobertura da terra.

A aplicação do método *RENN* para o tratamento de dados desbalanceados trouxe um aumento significativo no desempenho dos modelos, apresentando na aplicação dos modelos de predição, tanto utilizando o conjunto de dados de treinamento quanto de testes, um índice *Kappa* superior a 0.93.

Salienta-se que uma análise mais aprofundada nos métodos de tratamento de dados desbalanceados se fazem necessário, inclusive o uso conjunto de mais de uma solução, o que pode resultar em um menor tempo de processamento para o treinamento dos modelos e otimização dos hiperparâmetros.

Foi possível concluir que a utilização de plataforma e softwares livres gratuitos é uma abordagem aplicável para desenvolver modelos preditivos e sistemas de automatização de processos, possibilitando inclusive a reutilização dos códigos fontes para outros

objetos de classificação.

Os avanços na área de *machine learning* tem trazido não apenas algoritmos mais robustos e técnicas mais aprimoradas mas uma gama de plataformas como *Weka*¹⁵, *Orange*¹⁶ e *Knime*¹⁷ que facilitam o processo de implementação.

E apesar dos modelos de predição não terem sido colocados em produção e testados com outros conjuntos de dados, pelos resultados alcançados esta implementação se apresenta promissora.

Referências

- [1] M. C. Abreu, B. P. Conicelli, and J. R. Peñaranda, *Avaliação da Produtividade dos Poços Tubulares na Sub-Bacia do Juqueri-Cantareira/SP*, XIX Congresso Brasileiro de Águas Subterrâneas, (2016).
- [2] H. P. F. Alves, *Análise dos fatores associados às mudanças na Cobertura da Terra no Vale do Ribeira através da integração de dados censitários e de sensoriamento remoto*, fev. 2004. Tese, Departamento de Sociologia do Instituto de Filosofia e Ciências Humanas, Universidade Estadual de Campinas.
- [3] ANA, *Sala de Situação – Sistema Cantareira*, 2020. <https://www.ana.gov.br/sala-de-situacao/sistema-cantareira/>, acesso em: 03-09-2020.
- [4] G. E. A. P. A. Batista, *Pré-processamento de dados em aprendizado de máquina supervisionado*, 2003. Tese, Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo – USP, São Carlos – SP.
- [5] G. Borboudakis, T. Stergiannakos, M. Frysali, E. Klontzas, I. Tsamardinos, and G. E. Froudakis, *Chemically intuited, large-scale screening of MOFs by machine learning techniques*, npj Computational Materials, 3 (2017), pp. 1–7.
- [6] I. B. Carrijo, *Extração de regras operacionais ótimas de sistemas de distribuição de água através de algoritmos genéticos multiobjetivo e aprendizado de máquina*, 2004. Tese, Escola de Engenharia de Sã Carlos, Universidade de São Paulo, São Carlos – SP.
- [7] A. T. Caten, J. L. Safanelli, and L. F. C. Ruiz, *Mapeamento multitemporal da cobertura de terra, por meio de Árvore de decisão, na Bacia Hidrográfica do rio Marombas-SC*, Engenharia Agrícola, 35 (2015), pp. 1198–1209.
- [8] CPLA, *Mapeamento de Cobertura da Terra do Estado de São Paulo, 2010 – Escala 1:100.000*, 2013. <https://www.infraestruturaeambiente.sp.gov.br>, acesso em: 03-09-2020.
- [9] INPE, *Landsat*, 2020. <http://www.dgi.inpe.br/documentacao/>, acesso em: 03-09-2020.
- [10] M. Kubat, R. C. Holte, and S. Matwin, *Machine learning for the detection of oil spills in satellite radar images*, Machine learning, 30 (1998), pp. 195–215.
- [11] C. H. P. Luiz, *Modelagem da cobertura da terra e análise da influência do reflorestamento na transformação da paisagem: Bacia do Rio Piracicaba e Região Metropolitana do Vale do Aço*, mai. 2014. Dissertação, Programa de Pós-Graduação em Análise e Modelagem de Sistemas Ambientais, Universidade Federal de Minas Gerais, Belo Horizonte – MG.
- [12] E. L. Machado, *Um estudo de limpeza em base de dados desbalanceada e com sobreposição de classes*, 2007. Dissertação, Departamento de Ciências da Computação, Universidade de Brasília, Brasília – DF.
- [13] C. K. Matsukuma, *Análise comparativa de algoritmos de classificação digital não-supervisionada, no mapeamento do uso e cobertura do solo*, 2002. Dissertação, Universidade de São Paulo.
- [14] A. M. Neves, *Deteção remota de estruturas artificiais permanentes*, 2019. Dissertação, Faculdade

¹⁵<https://www.cs.waikato.ac.nz/ml/weka/>

¹⁶<https://orange.biolab.si/>

¹⁷<https://www.knime.com/>

de Ciências e Tecnologia, Universidade Nova de Lisboa.

- [15] A. R. S. Parmezan and G. E. A. P. A. Batista, *Descrição de modelos estatísticos e de aprendizado de máquina para predição de séries temporais.*, (2016). Relatório Técnico do ICMC, Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo, São Carlos – SP.
- [16] S. Paula, *Um estudo exploratório sobre a criação de modelos de predição cruzada de defeitos apoiada por uma medida de correlação*, 2016. Curso de Bacharelado em Ciência da Computação, Departamento Acadêmico de Computação, Universidade Tecnológica Federal do Paraná, Campo Mourão – PR.
- [17] J. L. Pereira, G. T. Batista, and D. A. Roberts, *Reflectância de Coberturas Vegetais na Amazônia*, Anais VIII Simpósio Brasileiro de Sensoriamento Remoto, Salvador, Brasil, (1996), pp. 551–556.
- [18] R. A. Petta, R. C. Fernandes, and P. S. R. Nascimento, *Detecção automática da dinâmica da cobertura da terra por sensoriamento remoto*, GEOGRAFIA (Londrina), 17 (2008), pp. 109–124.
- [19] F. A. Prado, *Sistema hierárquico de classificação para mapeamento da cobertura da terra nas escalas regional e urbana*, 2009. Dissertação, Programa de Pós-Graduação em Ciências Cartográficas, Universidade Estadual Paulista, Presidente Prudente – SP.
- [20] E. B. Ribeiro, *Classificação de Imagem Orbital Rapideye utilizando banco de dados NOSQL e método GEOBIA*, 2017. Dissertação, Programa de Pós-Graduação em Computação Aplicada, Universidade Estadual de Ponta Grossa, Ponta Grossa – PR.
- [21] J. Richards, *Remote sensing digital image analysis*, vol. 3, Springer, 1999.
- [22] H. G. Santos, *Comparação da performance de algoritmos de machine learning para a análise preditiva em saúde pública e medicina*, 2018. Tese, Programa de Pós-Graduação em Epidemiologia, Faculdade de Saúde Pública, Universidade de São Paulo, São Paulo – SP.
- [23] B. M. N. Souza, *Deteção e localização de fogo em imagens digitais usando técnicas de aprendizado de máquina*, 2019. Tese, Programa de Pós-Graduação em Informática, Pontifícia Universidade Católica do Paraná, Curitiba – PR.
- [24] USGS, *Landsat 5*, 2020. <https://www.usgs.gov/core-science-systems/nli/landsat/landsat-5>, acesso em: 03-09-2020.
- [25] G. M. C. Viglioni, *Metodologia para previsão de demanda ferroviária utilizando data mining*, 2007. Dissertação, Curso de Mestrado em Engenharia de Transporte, Instituto Militar de Engenharia, Rio de Janeiro – RJ.
- [26] T. Zimmermann, N. Nagappan, H. Gall, E. Giger, and B. Murphy, *Cross-project defect prediction: a large scale experiment on data vs. domain vs. process*, in Proceedings of the 7th joint meeting of the European software engineering conference and the ACM SIGSOFT symposium on The foundations of software engineering, 2009, pp. 91–100.