# CLASSIFICADOR SUPERVISIONADO RANDOM FOREST, APLICADO NA PLATAFORMA GOOGLE EARTH ENGINE.

Matheus Fillipe Gaia dos Santos [1]
Mauricio Alves da Silva [2]

## RESUMO

A produção de dados nunca alcançou patamares tão grandes quanto na história recente, porém, muitos dados estão desorganizados em meio a um "dilúvio" de informações criadas a cada segundo, e o grande desafio é encontrar maneira de extrair conhecimento. Partindo de tal ideia, a geografia faz o uso das geotecnologias para as tomadas de decisão técnicas, e a presente pesquisa usa desses princípios para apresentar como alternativa valida a programação em nuvem para o processamento de dados geográficos. A bacia hidrográfica do Rio do Coco, localizada ao oeste do estado do Tocantins e encontra seu exutório no Rio Araguaia. Por meio da plataforma de programação em nuvem *Google Earth Engine* – GEE, imagens Sentinel-2 para construção das imagens-base para classificação como composição colorida, índices espectrais NDVI e NDWI. Fazendo uso dessas bases, o classificador *Random Forest* – RF, através de 800 amostras, classificou a cobertura do solo nas classes de água, área descoberta, pastagem e formação floresta, resultando em acurácia geral do classificador 99,33%.

**Palavras-chave:** *Google Earth Engine*, *Random Forest*, Sentinel-2, Rio do Coco.

## ABSTRACT

In recent history, data production has reached unprecedented volumes, yet much of this data remains disorganized within an ever-expanding sea of information generated every second. The formidable challenge is to find ways to extract knowledge from this data deluge. Building upon this notion, geography leverages geotechnologies for technical decision-making, and this research explores cloud programming as a viable alternative for processing geographic data. Focusing on the Rio do Coco watershed, located in the western Tocantins state with its outlet in the Rio Araguaia, this study utilized Google Earth Engine (GEE) as a cloud-based programming platform. Sentinel-2 images were employed to create base images for classification, including false-color composites and spectral indices such as NDVI and NDWI. Using these bases, the Random Forest (RF) classifier, trained with 800 samples, successfully classified land cover into categories including water, bare land, pasture, and forest formation, achieving an impressive overall classifier accuracy of 99.33%. This research showcases the potential of cloud-based geoprocessing in the realm of geography and provides valuable insights into the management and utilization of vast geospatial datasets.

**Keywords:** Google Earth Engine, Random Forest, Sentinel-2, Rio do Coco.

[1]Mestrando do Programa de Pós-Graduação em Geografia da Universidade Federal de Santa Catarina – UFSC, santosgaia55@gmail.com;
[2]Professor dos Cursos de Geografia da Universidade Federal do Tocantins – UFT, mauricio.silva@mail.uft.edu.br;

## INTRODUÇÃO

O sensoriamento remoto de imagens orbitais oferece importante ferramenta para as análises espaciais, seja nas dimensões do tempo, do espectro eletromagnético e das formas que constituem determinado padrão no espaço. Possuindo as mais diversas aplicações, como os Modelos Digitais de Elevação – MDE e na aritmética bandas para criação de índices espectrais comumente usados como o *Normalized Difference Vegetation Index* – NDVI (Índice de Vegetação da Diferença Normalizada) ou *Normalized Difference Water Index* – NDWI (Índice de Água de Diferença Normalizada). Modelos de classificação supervisionada é apenas mais uma das ferramentas oferecidas pelo sensoriamento remoto e servem para diferenciar as diversar formas de apropiação sobre a cobertura solo, ou seja, diferenciando o seu uso.

O espaço geográfico possui diversas categorias para ser avaliado, como território, região, lugar e paisagem, porém, o modo de uso é algo transversal a todas as categorias citadas. Para Fitz (2017), os algoritmos de classificação servem para descriminar a cobertura e uso do solo, logo é necessário para sua aplicação, conhecer os parâmetros técnicos para formação de grupos de amostras, sendo eles o padrão das formas, tamanho, textura e contexto.

É necessário sinalizar que diferentes modelos matemáticos e estatísticos, resultam em generalizações diferentes, portanto é de máxima importância selecionar os com menores graus de confusão ou maior acurácia. Partindo dessa ideia Batista, (2022) testou vários modelos de Aprendizagem de Máquina para classificação da turbidez no Rio Paraobeba estado de Minas Gerais, sendo eles *Extra Trades* – ET, *Multilayer Perceptro* – MLP, *Naive Bayes* - NB, *Random Forest* – RF e *Support Vector Machine* – SVM, sobre o acervo de imagens Sentinel-2, e a pesquisa demonstrou que o RF e o SVM são aqueles com os melhores resultados de acurácia do geral classificador superior a 85% obtidos pelos conjuntos de amostras.

O *Random Forest* é fundamentado nos conceitos de aleatoriedade e majoritariedade, ou seja, os conjuntos construídos a partir de regras, mas as amostras são aleatórias, junto disso a decisão majoritária é feita a partir do resultado de várias árvores decisão, na qual os resultados da maioria determinam a classe o píxel em uma votação chamada de *Bagging*. Por isso, algoritmos de agrupamento com o *Random Forest* aumentam a precisão por meio da combinação de resultados. (BRUCE; BRUCE, 2019).

Parte importante das rotinas de sensoriamento remoto e geoprocessamento, é a escolha da plataforma para processamento das imagens, O *Google Earth Engine* – GEE é uma plataforma de processamento em nuvem de dados geográficos. E segundo Souza, Moreira e Machado (2009), a programação em nuvem agiliza o processamento para grandes volumes de

informações em plataformas computacionais de terceiros, e isso nasce da necessidade dos usuários consultar, construir e configurar sistemas virtuais, demonstrando o principal intuito do processamento em nuvem é reduzir drasticamente o tempo de resposta dos comandos e das ações feitas em ambiente computacional. Tendo em vista esses cenários que exigem o processamento acelerado de dados, a programação em nuvem junta-se às geotecnologias para o estudo dos fenômenos naturais e resolução dos conflitos socioambientais.

O GEE é uma plataforma de sensoriamento remoto e processamento de imagem em escala planetária com 1 quatrilhão de dados codificado em *JavaScript*, de uso gratuito que permite o usuário ter acesso ao banco de imagens das mais diversas séries de satélites como, Landsat, Modis, NOAA, Sentinel e outros, o mesmo vale para os dados vetoriais que também contam com uma ampla variedade.

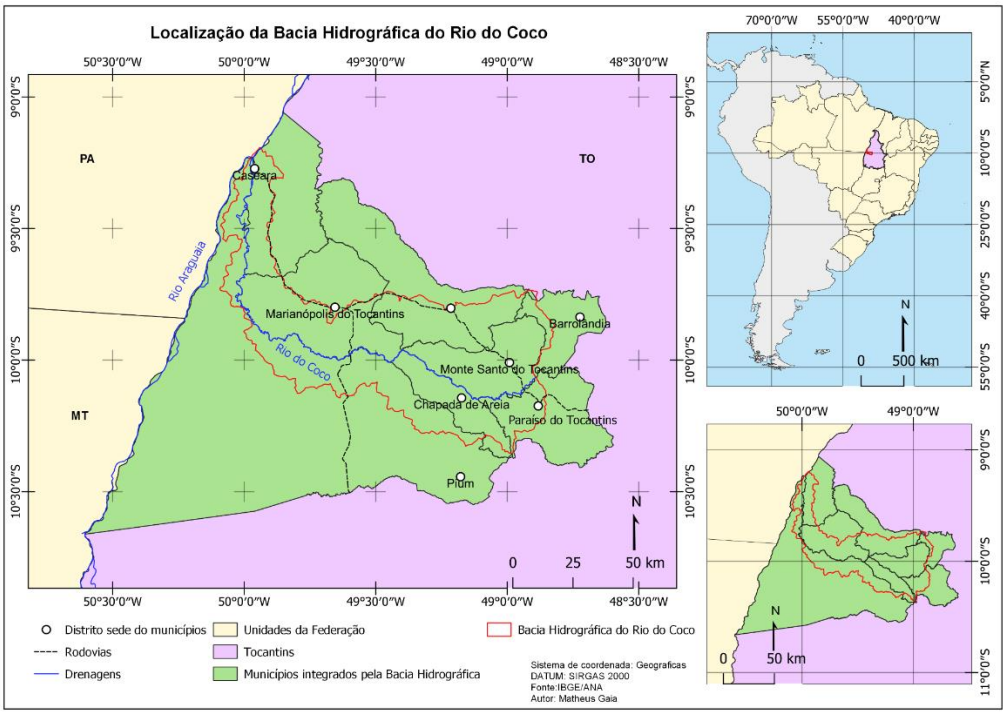
O presente trabalho utiliza da Bacia Hidrográfica do Rio do Coco, localizada no estado do Tocantins, usando imagens do satélite Sentinel-2 no ano de 2020, para aplicação do algoritmo de classificação utilizando o modelo de Aprendizagem Maquina Random Forest no GEE.

## METODOLOGIA

A Bacia Hidrográfica do Rio Coco conta com uma área de 6,670 Km², com o curso de água principal, o Rio do Coco medindo 356 Km de comprimento, suas principais nascentes estão na Serra do Estrondo e a foz encontra-se no Rio Araguaia representada na figura (1).

A sua dimensão ocupa parcialmente vários municípios, seja no alto ou baixo curso da bacia hidrográfica, como Barrolândia, Monte Santo do Tocantins, Chapada de Areia, Paraíso do Tocantins, Pium, Marinópolis do Tocantins e Caseara. Que são acessados pelas as rodovias estaduais, TO-080 que corta do alto ao baixo Rio do Coco na direção leste e oeste, já a TO-374 corta a área de estudo na direção Norte e Sul.

Figura 1: Mapa de localização da área de estudo




Fonte: Autores, (2022).

As imagens selecionas foram obtidas na coleção de imagens do satélite Sentinel-2 sensor *MultiSpectral Instrument* – MSI de seu uso gratuito e resolução espacial de 10 metros, foram usadas as bandas que correspondem as faixas do espectro eletromagnético B4 (vermelho), B3 (verde), B2 (azul) e B8 (infra-vemelho próximo), junto disso foi aplicada a aritmética bandas para construir os índices normalizados NDVI e NDWI. É preciso ressaltar, que parte do trabalho foi dedicado a seleção imagens com mínimos efeitos atmosféricos, visando diminuir ruídos e confusões no resultado, sendo assim as imagens que são obtidas correspondem aos meses de agosto até outubro de 2020, porque tais meses representam a estação seca.

A Construção de dados *raster* no GEE é resultado de uma coleção de imagens, definida de acordo com periodização da área de estudo, porém as inúmeras imagens da coleção precisam de um redutor matemático para a aplicação das próximas operações feitas no *script*, o redutor usa a operação de mediana para selecionar imagens pouco influenciadas por valores extremos de refletância, visando reduzir os erros de classificação.

Portanto, o arranjo de imagens usadas para classificação envolveu a composição colorida do Sentinel-2 com bandas B4, B3 e B2, aliado os índices espectrais NDVI construídos a partir das Bandas B8 e B4, e por conseguinte o NDWI usa as bandas B8 e B3. O algoritmo

criado em *JavaScript* no GEE possui etapas, ou seja, definições de parâmetros para seleção das imagens (filtro de nuvens), delimitação da área classificada (polígono com os limites da bacia hidrográfica), criação da variável que armazena as imagens do Sentinel-2 e das respectivas bandas citadas conforme o período temporal de agosto até outubro de 2020, aplicação dos índices NDVI e NDWI, criação dos grupos de amostra, num total de 4 classes, sendo de água, área descoberta, formação campestre e formação florestal. A etapa de aplicação do modelo de Aprendizagem de Máquina RF que é responsável pela união do conjunto de amostras com os grupos de composição colorida (B4, B3 e B2), NDVI e NDWI. E por fim a exibição dos valores de matriz de confusão, acurácia geral, acurácia do usuário e acurácia do produtor.

Para maximizar os melhores resultados e diminuir a confusão por parte do algoritmo classificador RF, as imagens do Sentinel-2 do nível 2A *Surface Reflectance* – SR foram usadas devido aos processos avançados de correção radiométrica e atmosférica, que resultam em medidas de refletância mais acuradas (ESA, 2015).

A calibração das amostras usadas pelo RF, consistem em definir os mesmos limites e condições para os conjuntos de amostras de cada classe. Foi definido 200 amostras para cada grupo totalizando 800, na qual 70% das amostras foi destinada para conjunto de treinamento, isto é, o que de fato é usado para alimentar o classificador RF, outra parte importante para validação do classificador é o conjunto de teste que possui 30% das amostras, necessário para confrontar os dados de treinamento.

A proporção dos conjuntos para os conjuntos de treinamento e de teste, seguem a proporção de distribuição gaussiana conforme o teorema do limite central, se as amostras de uma população são iguais ou maiores que 30, a distribuição das médias tende a se aproximar de uma distribuição normal (MONTGOMERY; RUNGER, 2003).

Para auxiliar na temática de cobertura do solo, foi usado a paleta de cores e instruções para a construção dos mapas segundo IBGE (2013), que lança uma série de diretrizes para o mapeamento do solo brasileiro conforme as novas referências técnicas, que dependem da aplicação de geotecnologias.

## REFERENCIAL TEÓRICO

De acordo com Florenzano (2012), o Sensoriamento Remoto é a técnica que consiste na captação e registro a distância, e estritamente sem contato direto com a fonte de energia, e

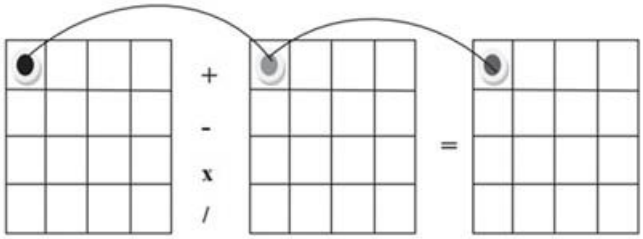
sensores são os equipamentos capazes de receber ou emitir energia a fim de registrá-la em formato digital, e para posteriormente serem editados, filtrados e analisados por meio de softwares específicos.

A classificação supervisionada feita na pesquisa se apropria de uma das áreas do sensoriamento remoto chamada dimensão do espectro, é o modo de identificação e diferenciação dos alvos através da resposta espectral. A título de exemplo, os corpos hídricos tendem a retornar para o sensor uma refletância diferente da vegetação ou do solo, e isso deve a maneira que cada porção do espectro eletromagnético representado pelas bandas que interage com os alvos. Então, a seleção dos alvos precisa seguir critérios técnicos para formação do grupamento e amostras que representam determinado tipo de alvo.

Os elementos da interpretação de imagens orbitais, podem ser traduzidos com a caracterização das superfícies por meio de critérios que obedecem a um conjunto fatores fisiográficos e antrópicos. Começando com padrão de formas, ou seja, os elementos geométricos que tendem a ser associados lavouras e edificações, ou alvos com padrões menos regulares como os rios. O tamanho, é outro elemento importante, pois ele revela a escala e extensão dos fenômenos, por exemplo, um loteamento residencial ou fenômenos naturais como impacto socioambiental como deslizamentos. A textura, que está relacionada com a escala e sua capacidade de mostrar diferentes agrupamentos que provocam variações nos tons do alvo, exemplo da textura áspera geralmente atribuídas as florestas heterogêneas ou grosseiras relacionadas com os relevos acidentados. As características que estão juntas dos elementos da interpretação, são as condições referente ao contexto do terreno. Primeira dessas características é localização, portanto, é algo relacionado com as adjacências do alvo. O brilho ou a tonalidade, é alvo diz respeito as propriedades físicas do alvo e suas condições superficiais. (FITZ, 2017).

A aritmética de bandas/imagens de sensoriamento remoto é algo aplicado píxel a píxel, visando produzir realces sob determinadas características e alvos por meio de operações matemáticas envolvendo duas ou mais imagens figura (2). Logo, o NDVI representado pela equação NDVI = ρnir–ρr/ρnir+ρr, em que (ρnir) é a banda do infravermelho próximo e (ρr) representa a banda do vermelho, já para o NDWI a equação NDWI= ρnir-ρg/ρnir+ρg que também usa o infravermelho próximo junta ao (ρg) que é a banda do verde. Sendo o NDVI importante para identificação do vigor vegetacional, portanto, útil para diferenciar diferentes formações vegetais, aliando ao NDWI usado para destacar corpos hídricos e a concentração de água nas estruturas vegetais, para ambos as variações são de -1 até 1. (MENESES; ALMEIDA, 2012).

Figura 2: Aritmética de Bandas




Fonte: Meneses e Almeida, (2012).

De acordo com Tagliarini, et al (2017) em Botucatu estado de São Paulo, o NDVI e NDWI foram aplicados usando bandas do satélite LANDSAT-8, com o intuito de construir mapas de classificação supervisionada usando composição colorida 4,3 e 2 junto aos índices espectrais na foi obtido ótimos resultados devido à sensibilidade dos índices.

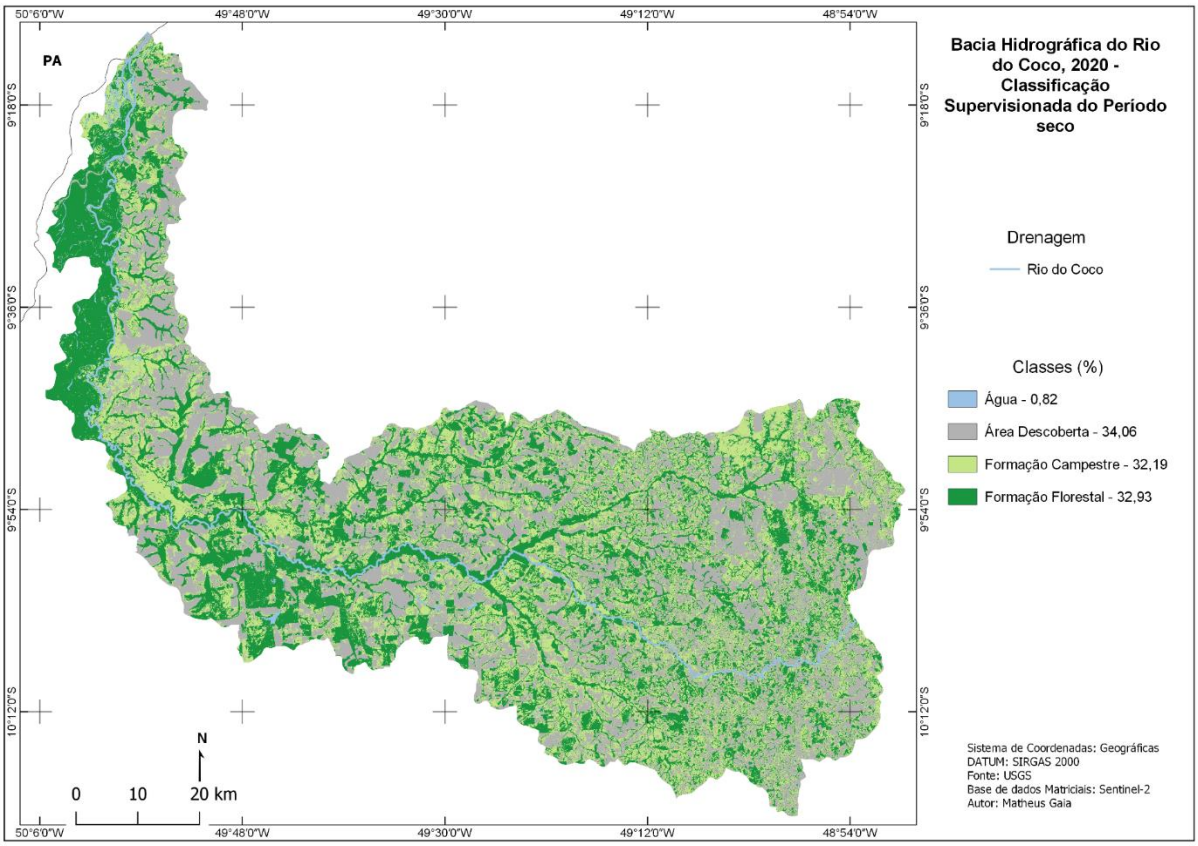
O processamento paralelo, fundamentado pela estrutura Servidor-Cliente disponibilizado pelo GEE, possibilitou uma equipe de pesquisadores do Rio Grande do Sul, a análise de uso e cobertura solo junto a variações pluviométricas em períodos sazonais. Pois a plataforma possibilita o processamento de dados nativos GEE, aliado ao acervo de dados disponibilizado pela equipe da pesquisa. (MATEUS. et al, 2021).

## RESULTADOS E DISCUSSÃO

A criação do algoritmo no *JavaScript* consiste em estruturar uma série de ordens hierárquicas, ou seja, aquelas que nas linhas superiores possuem prioridades, pois as informações armazenadas nas variáveis servem para as modificações das variáveis posteriores. A exemplo, a função que filtrar as cenas sem nuvens é algo precisas ser definido antes da variável que armazenas as cenas usadas na classificação. Logo, outro princípio importante é a lógica, os comandados não podem contradizer as ordens hierárquicas definidas anteriormente.

Já estabelecido todos os parâmetros para o RF, a classificação foi obtida com os seguintes resultados para as 4 classes de acordo com a figura (3). A água constitui 0,82% que corresponde a uma área de 5,484ha, respectivamente a área descoberta abrange 228,345ha ou 34,06%, para a classe formação campestre ocupa 215,74ha que proporcionalmente é 32,18%, já a formação florestal possui 220,759ha ou 32,93%.

Figura 3: Mapa de classificação supervisionada




Fonte: Autores, (2022).

Para validar tais resultados, o GEE disponibiliza para o usuário a matriz de confusão tabela (1), que nada mais que um dado tabular formado pelos conjuntos de amostra feito pelo usuário (linhas), e conjunto de amostras testadas pela modelo classificador (colunas). Com essa matriz em mãos, foi calculado a acurácia geral do classificador que resultou em 99,33%. Posteriormente, os erros de comissão e omissão são demonstrados como acurácia do usuário e acurácia do produtor que variam de 0 até 1, quanto mais o valor tende ao zero maior é seu erro.

Tabela 1: Matriz de confusão

| Matriz de confusão | | | | | | |
|---|---|---|---|---|---|---|
| | Não classificado | Água | Solo Exposto | Pastegem | Formação Florestal | Total de amostras |
| Não classificado | 0 | 0 | 0 | 0 | 0 | 0 |
| Água | 0 | 57 | 0 | 0 | 0 | 57 |
| Solo Exposto | 0 | 0 | 108 | 1 | 0 | 109 |
| Pastegem | 0 | 0 | 0 | 65 | 1 | 66 |
| Formação Florestal | 0 | 0 | 0 | 0 | 69 | 69 |
| Total | 0 | 57 | 108 | 66 | 70 | 301 |

Fonte: Autores, (2022).

A acurácia do usuário, apresenta as imprecisões de inclusão de determinado píxel em uma classe na qual ele não pertence, por isso também é chamado de erro de comissão, e nesta classificação ocorre os valores de 0,98 foram demonstrados nas classes de formação campestre e formação florestal, o restante das classes ficou com valor 1. Já a acurácia do produtor, sinalizar os erros de omissão, ou seja, a exclusão do píxel de uma área na qual ele pertence, neste caso as classes de área descoberta e a formação campestre com valor de 0,98, enquanto as outras duas classes permaneceram com 1. O resultado desses dois parâmetros explicita na tabela (2) o desempenho individual das classes, desse modo é possível compreender com os erros de determinada parte repercutem no mapa.

Tabela 2: Matriz de confusão

| Acurácia do usuário | valor | Acurácia do produtor | valor |
|---|---|---|---|
| Não classificado | 0 | Não classificado | 0 |
| Água | 1 | Água | 1 |
| Solo Exposto | 1 | Solo Exposto | 0,990825688 |
| Pastegem | 0,984848485 | Pastegem | 0,984848485 |
| Formação Florestal | 0,985714286 | Formação Florestal | 1 |

Fonte: Autores, (2022).

Sendo assim, os valores próximos a 0,98, indicam erros pouco significativos, seja de inclusão ou exclusão, demonstrando grande acurácia das tomadas de decisão feitas pelo RF. Tais erros estão associados as semelhanças espectrais dos alvos, pois a classificação supervisionada de modelo é limitada ao reconhecimento de diferenças espectrais.

## CONSIDERAÇÕES FINAIS

A criação de um algoritmo que usa como base o modelo de Aprendizagem de Máquina, o *Random Forest*, com uma acurácia geral de 99,33%. No entanto, acurácia geral classificador não é o mesmo que acurácia global do mapa, pois aquilo que foi verificado diz respeito aos grupos de amostras definidos unicamente em ambiente virtual, fica evidente que os dados que alimentaram o classificador passaram por seleção criteriosa. Porém, a exatidão global do mapa é feita pelo trabalho e verificação em campo usando pontos de controle.

Feito esse adendo os resultados obtidos no ambiente programação em nuvem GEE são satisfatórios, muito disso deve ao processo ágil, aliado a um arsenal imenso de ajustes que podem ser feitos pelo pesquisador. Os algoritmos estão ganhando cada vez mais espaço nos trabalhos que envolvem sensoriamento remoto, logo é importante testar outros modelos de Aprendizagem Maquina em ambientes de programação em nuvem.

## REFERÊNCIAS

BATISTA, L. V. Turbidity Classification of the Paraopeba River Using Machine Learning and Sentinel-2 imagens. **IEE Latin America Transaction**, New York: v.20, n.5, p.799-805. 2022.

BRUCE, P; BRUCE, A. Aprendizado de Máquina Estatística. In:_____. **Estatística Prática para Cientistas de Dados**: 50 Conceitos Essenciais. 1ed, Rio de Janeiro: Alta Books, 2019. p.215-255.

FITZ, P. R. Sensoriamento Remoto e Sistemas de Informação Geográfica. In:_____. **Geoprocessamento Sem Complicação**. 4 ed. São Paulo: Oficina de Textos, 2017. p.97-138.

ESA (European Space Agency). **Sentinel-2 User Handbook**. 2 ed. London: ESA Standard Document, 2015.

FLORENZANO, T. G. **Geomorfologia**: conceitos e tecnologias. 5 ed. São Paulo: Oficina de Textos, 2008.

MATEUS, M. G. et al. Visualizador de Água e Solo: Uma aplicação voltada para o gerenciamento de recursos naturais desenvolvida na plataforma Google Earth Engine. In: WORKSHOP DE COMPUTAÇÃO APLICADA À GESTÃO DO MEIO AMBIENTE E RECURSOS NATURAIS (WCAMA), 12. 2021, Evento Online. **Anais** [...] Porto Alegre: Sociedade Brasileira de Computação, 2021. p. 157-166.

MENESES, P. R; ALMEIDA, T. Aritmética de Bandas. In:_____. **Introdução ao Processamento de Imagens de Sensoriamento Remoto**. 1 ed, Brasília: UnB. 2012. p.138-151.

MONTGOMERY, D. C; RUNGER, G. C. Estimação de parâmetros. In:_____. **Estatística Aplicada e Probabilidade para Engenheiros**. 2 ed. Rio de Janeiro: Livros Técnicos e Científico Editora, 2003. p. 128-141.

SOUSA, F; MOREIRA, L; MACHADO, J. Computação em nuvem: conceitos, tecnologias, aplicações e desafios. In: III Escola Regional de Computação Ceará, Maranhão, Piauí - ERCEMAPI, 2009, Parnaíba-PI. **Anais** [...] ERCEMAPI, Parnaíba: SBC, 2009.

TAGLIARINI, F. S. N. et al. Índices NDVI e NDWI como ferramentas ao mapeamento do uso e ocupação do solo em bacia hidrográfica. In: Simpósio Brasileiro de Sensoriamento Remoto-SBSR, 18. 2017, Santos. **Anais**[...] Santos: INPE, 2017, p.2271-2278.