Random Forest-Based Land Use and Land Cover Classification in Brazil's Cerrado

Khalil Ali Ganem – University of California, Los Angeles (UCLA)

Abstract: *This project, conducted as the final requirement for the A&OS C204 course at the University of California, Los Angeles (UCLA), evaluates the effectiveness of the Random Forest (RF) algorithm for land use and land cover (LULC) classification in São Desidério, a deforestation hotspot in Brazil's Cerrado ecoregion. Leveraging cloud-free Sentinel imagery, RF was chosen for its ability to manage spectrally similar classes and resist overfitting. Third-party training data were used to classify cultivated land and natural vegetation, providing insights into the region's ecological status. The results revealed that nearly half of São Desidério's native vegetation has been lost due to agricultural expansion. The classifier achieved a test accuracy of 90% and training accuracy of 100%, effectively differentiating between natural and anthropogenic land cover types. Incorporating near infrared and shortwave infrared bands significantly enhanced classification accuracy, underscoring their importance in LULC studies. This work demonstrates the reliability of RF for LULC classification, supporting conservation and monitoring efforts.*

1. Introduction

The Cerrado is a Brazilian ecoregion comprised of a mosaic of native grasslands, savannas, and forests, and one of the most biodiverse savannas in the world. However, 46% of its original cover has been cleared to make way for crops and pastures, paving the way for a new agricultural frontier, responsible for 12% of global soybean production (Russo et al., 2018) and 10% of global beef exports (Organisation for Economic Co-operation and Development, & Food and Agriculture Organization & Food and Agriculture Organization of the United Nations, 2021). São Desidério is one of the areas most affected by this rapid expansion and has been pointed out as Brazil's new deforestation hotspot (Araújo et al., 2019), highlighting how much attention needs to be devoted to understanding how much native vegetation is left untouched in the city's area to enhance monitoring efforts towards preserving local biodiversity.

To achieve this, refined technology is necessary, and Remote Sensing (RS) becomes essential. RS is the science of obtaining information about objects or areas from a distance, typically using satellite or airborne sensors to collect data across a range of wavelengths (Fischer et al., 1976). It allows for comprehensive monitoring of large and inaccessible areas, enabling researchers to assess vegetation, land cover, and other critical environmental variables over time. One of the most effective applications of RS to enhance conservation practices is the development of land use/cover (LULC) maps, where pixels in an image are classified based on their spectral characteristics. To create an accurate LULC map, three key elements are required: high-resolution, cloud-free satellite imagery; reliable training and validation data to classify LULC accurately and assess the reliability of the final product; and a robust machine learning-based approach to effectively handle the classification. Machine learning (ML), first defined by Arthur Samuel in 1959, refers to the ability of computers to learn and improve from experience without being explicitly programmed (Samuel, 1959). It has been employed in numerous applications, including natural language processing, predictive modeling, medical diagnostics, and environmental monitoring, as well as for classifying satellite imagery into useful categories such as forests, and croplands. Therefore, this project's primary goal is to classify LULC in São Desidério using RS employing a supervised machine-learning approach. The expected result is a reliable LULC map with a classifier's reported accuracy of at least 80% and statistics on current LULC distribution. This project will test the potential of ML-based supervised classification for LULC assessment, justifying its application in other Brazilian municipalities impacted by natural and anthropogenic changes.

2. Data

To generate an image mosaic, I utilized Google Earth Engine (GEE) and Google Colab, enabling access to cloud-free, high-quality datasets efficiently. GEE's vast catalog and computational capabilities facilitate processing and analysis of Earth observation data (Gorelick et al., 2017), including Sentinel-2 imagery, which was selected for this study due to its high spatial and spectral resolution (Figure 1). Sentinel-2 is a wide-swath, multi-spectral imaging mission supporting the Copernicus Land Monitoring Program. It offers spatial resolutions (i.e., pixel size) of 10, 20, and 60 meters across 13 spectral bands, covering the visible, red edge, near infrared, and shortwave infrared spectra (Drusch et al., 2012). These characteristics make Sentinel-2 well-suited for distinguishing between natural and anthropogenic areas. The dataset is available in the Google Earth Engine catalog.

Figure 1 – 2018 Sentinel-2 mosaic of São Desidério (15,157 km²) with enhanced Red-Green-Blue visualization (i.e., true-color) using 100% linear stretch.

To prevent bias during training and validation, I incorporated third-party data from a study conducted by researchers at Brazil's National Institute for Space Research (Ferreira et al., 2020). This dataset, accessible here, includes samples collected through visual interpretation of high-resolution imagery for different years. The samples are classified into four categories: Cropland, Pastureland, Cerrado, and Cerradão, with the latter two representing natural vegetation. By using this independent dataset, which corresponds to the same year as the Sentinel-2 image (2018), I reduced the risk of introducing bias during the training and validation process. This approach ensures that the classification model is evaluated against a dataset not derived from the model itself, enhancing the reliability and accuracy of the classification.

3. Modelling

I adopted a systematic approach that integrates data preprocessing, feature extraction, and machine learning modeling aiming at achieving high classification accuracy. I installed all necessary Python packages in Google Colab and authenticated and initialized the Earth Engine API to access Sentinel-2 imagery for São Desidério. I defined the study area using a pre-existing shapefile of São Desidério, imported into GEE as a *FeatureCollection.* To ensure spatial consistency during processing, I converted the shapefile into a polygon and integrated a labeled dataset of geographic sample points with land cover categories. These points were clipped to the study region using *geopandas*, and their categorical labels were transformed into numeric codes because raster data, which represents images as grids of pixels, requires numerical values for processing and analysis in machine learning workflows. Additionally, I resampled the points to align with the pixel size of the Sentinel-2 mosaic, ensuring accurate correspondence between the labeled data and the image grid. Using *geemap*, I selected the spectral bands (Table 1) and excluded those prone to errors (e.g., probability bands). Third-party training data, originally stored in *.rda* format (specific to R), required integration between R and Python because the data was created and saved using R-based tools. To use this data in the Python-based ML workflow, I processed it using *rpy2*, a library that allows seamless interaction between R and Python, enabling me to load the R data directly, extract the labeled information, and convert it into a structured format suitable for Python-based analysis. To simplify the analysis and enhance interpretability, I harmonized the labels into broader categories, such as "Cultivated Land" and "Natural Vegetation" (Figure 2).

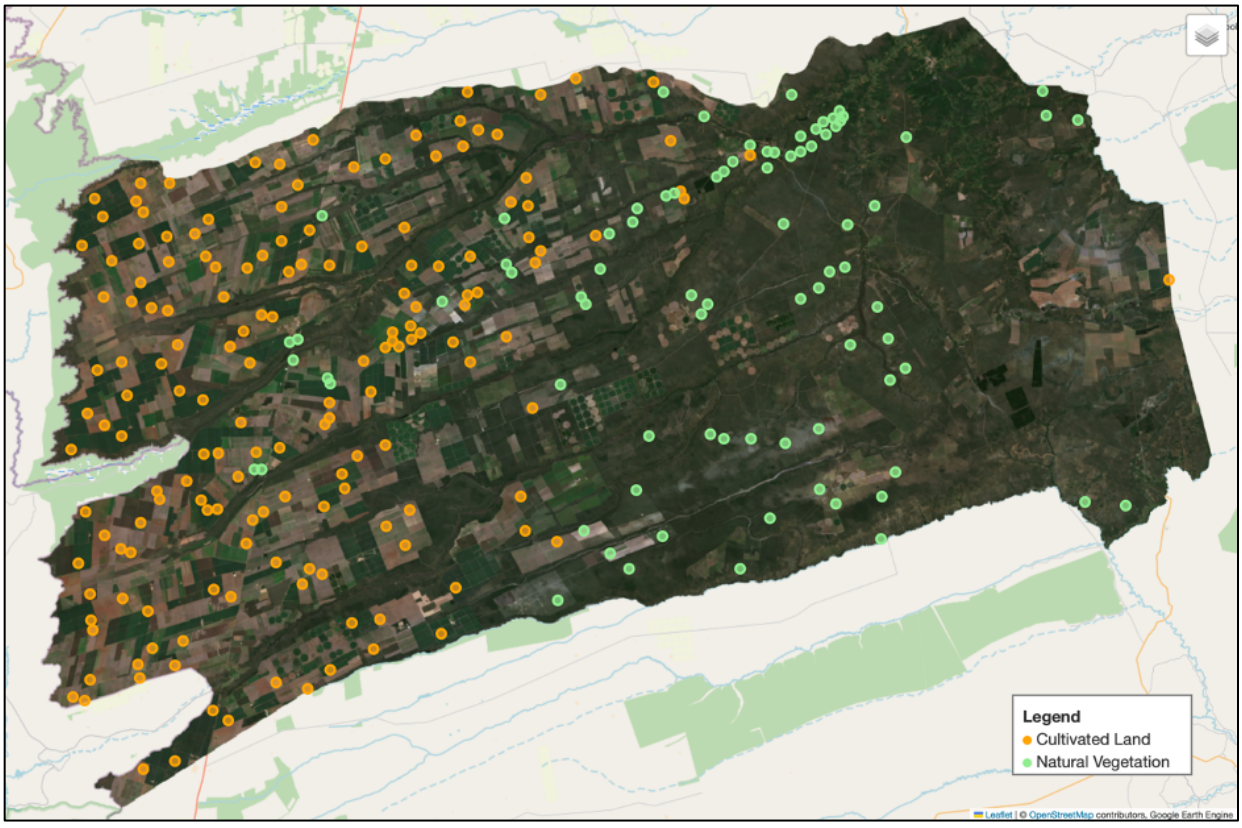

Figure 2 –Training/Validation samples (246) overlaid in Sentinel-2 mosaic of São Desidério for both Cultivated Land (163) and Natural Vegetation (83).

For classification, I used a Random Forest (RF) algorithm from *sklearn*, which I selected for its ability to handle high-dimensional data and its robustness against overfitting. Literature on LULC classification has pointed out the outperformance of RF comparing to other supervised classifiers like artificial neural networks, maximum likelihood, and support vector machines in accuracy (Talukdar et al., 2020). RF, an ensemble machine-learning algorithm, enhances prediction by building multiple decision trees (Breiman, 2001). It is non-parametric, robust to outliers, and effective in noisy data environments. Each decision tree predicts independently, and the final prediction is the majority vote across all trees (Gislason et al., 2006). This ensemble approach improves model accuracy and stability. I configured the RF model with 100 decision trees (n_estimators=100) and set a fixed random state (random_state=42) to ensure reproducibility. I split the dataset into training and testing subsets in an 80:20 ratio using train_test_split. The feature set for the classifier includes all bands of Table 1 in addition to Aerosol Optical Thickness (AOT), True Color Image - Red (TCI_R), True Color Image - Green (TCI_G), True Color Image - Blue (TCI_B), Scene Classification Layer (SCL), and Water Vapor (WVP).

Table 1 - List of spectral bands and their specifications

Band Name	Band Number	Wavelength (nm)	Spatial Resolution (m)
Coastal Aerosol	B1	443	60
Blue	B2	490	10
Green	B3	560	10
Red	B4	665	10
Red Edge 1	B5	705	20
Red Edge 2	B6	740	20
Red Edge 3	B7	783	20
Near-Infrared (NIR)	B8	842	10
Narrow NIR	B8a	865	20
Water Vapor	B9	945	60
Shortwave Infrared 1 (SWIR)	B10	1375	60
Shortwave Infrared 2 (SWIR)	B11	1610	20
Shortwave Infrared 3 (SWIR)	B12	2190	20

4. Results & Discussion

I applied predictions to the entire Sentinel-2 image, generating a map classified into “Cultivated Land” and “Natural Vegetation” (Figure 3). The model achieved a test accuracy of 90% and a training accuracy of 100%, demonstrating an excellent fit to the training data while maintaining strong generalization to unseen data. The model produced clear results, with cultivated areas in the central portion distinctly delineated, particularly the characteristic circular crop patterns formed by center-pivot irrigation systems. The well-defined farmland boundaries also emphasize the model's ability to capture the geometric imprint of human activity. This highlights the model's effectiveness not only quantitatively but also from a qualitative perspective, as the classification accurately captures observable land use patterns.

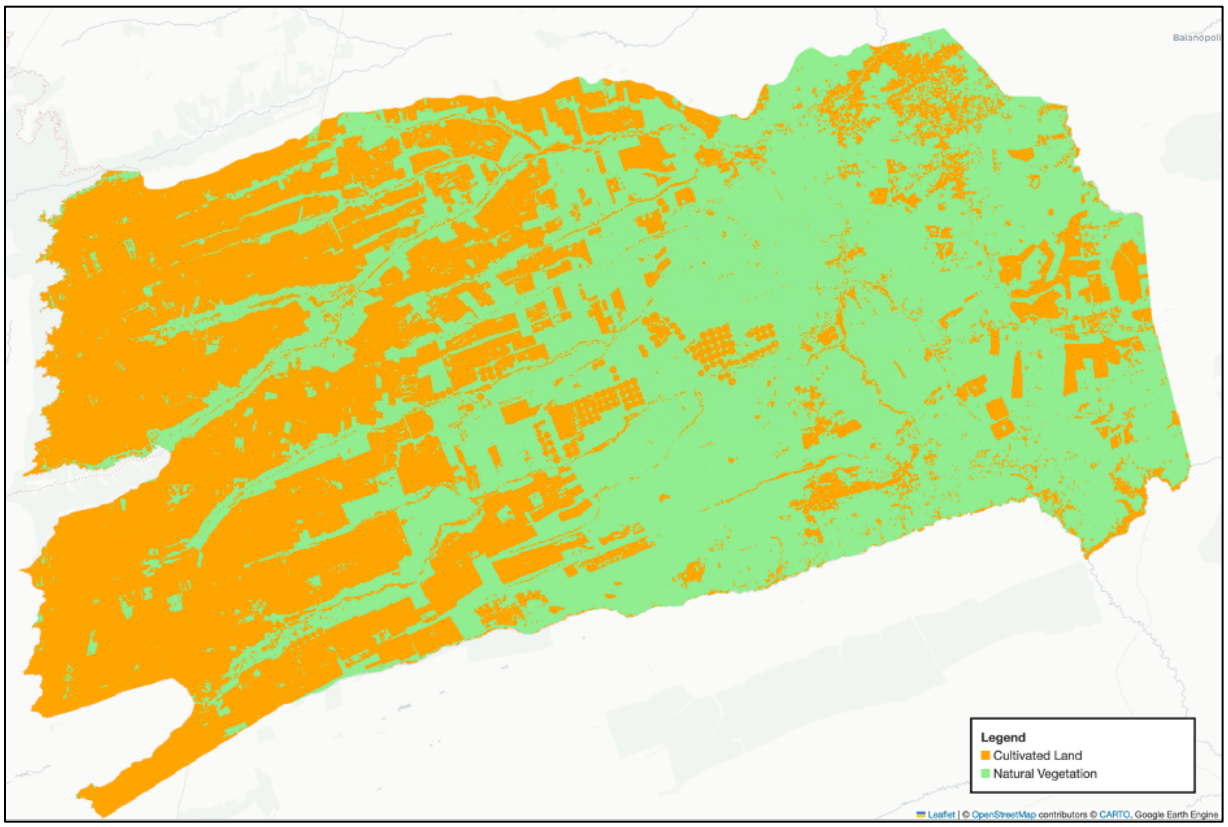

Figure 3 – 2018 LULC classification result of São Desidério.

I evaluated the RF classifier's performance by generating a classification report with *sklearn* and using metrics such precision (the proportion of true positives out of all predicted positives), recall (the proportion of true positives out of all actual positives), F1-score (the harmonic mean of precision and recall, balancing false positives and false negatives), and support (the total number of actual instances for each class).

The classification results highlight strong model performance for both land cover classes, with "Cultivated Land" achieving a precision of 0.89, recall of 0.97, and F1-score of 0.93, and "Natural Vegetation" achieving a precision of 0.93, recall of 0.76, and F1-score of 0.84 (Table 2). The high precision for both classes indicates the model makes few false positive errors, while the higher recall for "Cultivated Land" suggests the model effectively captures most of these areas, likely due to the distinct spectral characteristics of croplands and pasturelands.

Conversely, the lower recall for "Natural Vegetation" indicates some under-detection, likely caused by spectral overlap with "Cultivated Land", particularly during peak photosynthetic activity when cropland exhibits similar reflectance patterns across most bands, leading to potential confusion (Shao et al., 2010). This issue is intensified by the dataset's imbalance, with 163 samples for "Cultivated Land" compared to 83 for "Natural Vegetation", potentially biasing the model toward better recognition of the majority class. While the results demonstrate the model's overall robustness, the slight bias toward over-detecting "Cultivated Land" underscores the need for improvements in balancing precision and recall for "Natural Vegetation". Addressing this imbalance through techniques such as oversampling, under sampling, or enhanced feature engineering could lead to more equitable performance across both classes.

Table 2 – Classification report generated by *sklearn*

Class	Precision	Recall	F1-Score	Support
Cultivated Land	0.89	0.97	0.93	33
Natural Vegetation	0.93	0.76	0.84	17

To assess the spatial distribution of classified land cover types, I calculated the pixel-based proportions for each class (Figure 4). The results show that "Cultivated Land" accounts for 46.7% of the area, while "Natural Vegetation" comprises 53.3%. The findings align with established land use practices and patterns of deforestation and vegetation replacement, confirming São Desidério's status as a deforestation hotspot. Additionally, the clear west-to-east deforestation trend observed in the classification (Figure 3) corresponds to similar patterns reported in other parts of the Cerrado, further validating the results (Almeida De Souza et al., 2020).

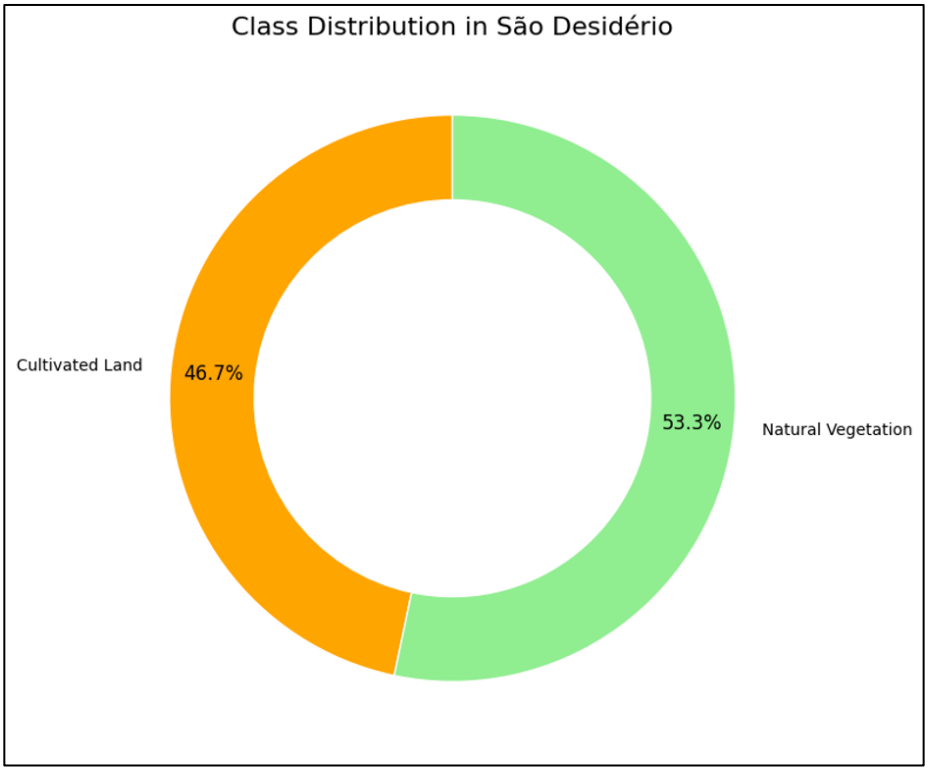

Figure 4 - Class Distribution in São Desidério (based on Figure 3).

To understand the model's decision-making process, I computed feature importance, identifying which spectral bands (variables) contributed most to the classification model. The feature importance values from the RF model were normalized to a 0 – 1 scale to ensure consistent comparability across features. Without normalization, feature importance values are relative and can be compared but lack a uniform scale, making interpretation less intuitive. I also categorized the normalized values into four color-coded ranges to enhance interpretability (Figure 5). Shortwave Infrared (B12 and B11) and Near-Infrared (B8) bands dominate the model's performance, confirming their critical role in distinguishing vegetation and soil properties. Their high importance highlights their sensitivity to moisture and biomass, essential for separating "Cultivated Land" from "Natural Vegetation". In contrast, visible spectrum bands (B2, B3, and

TCI_B) contribute minimally, emphasizing their limited utility in this classification task. The moderate relevance of B4 and B9 suggests these bands provide complementary information.

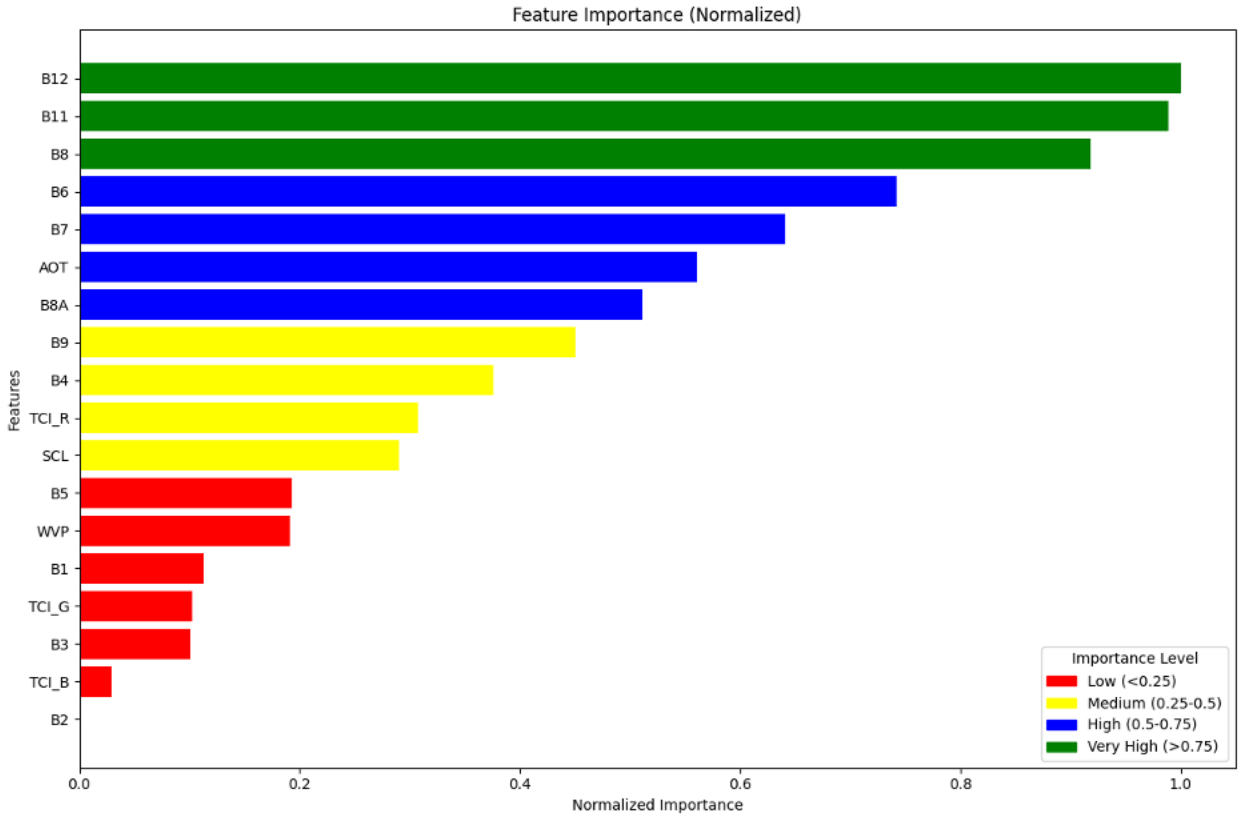

Figure 5 - Random Forest Variable Importance (Normalized) and categorized by different levels of importance

To further evaluate the model's performance, I used Receiver Operating Characteristic (ROC) curves to assess its ability to separate the classes effectively (Figure 6). The ROC curve illustrates the model's ability to distinguish between the two land cover classes by plotting the true positive rate (sensitivity) against the false positive rate (1-specificity) at various classification thresholds. The AUC (area under the curve) of 0.94 reflects excellent model performance, indicating that the classifier correctly discriminates between "Cultivated Land" and "Natural Vegetation" with a high degree of accuracy. The curve's proximity to the top-left corner demonstrates the model's effectiveness in minimizing false positives while maximizing true positives. In contrast, the dashed diagonal line represents a random classifier, emphasizing the superiority of the model's predictions.

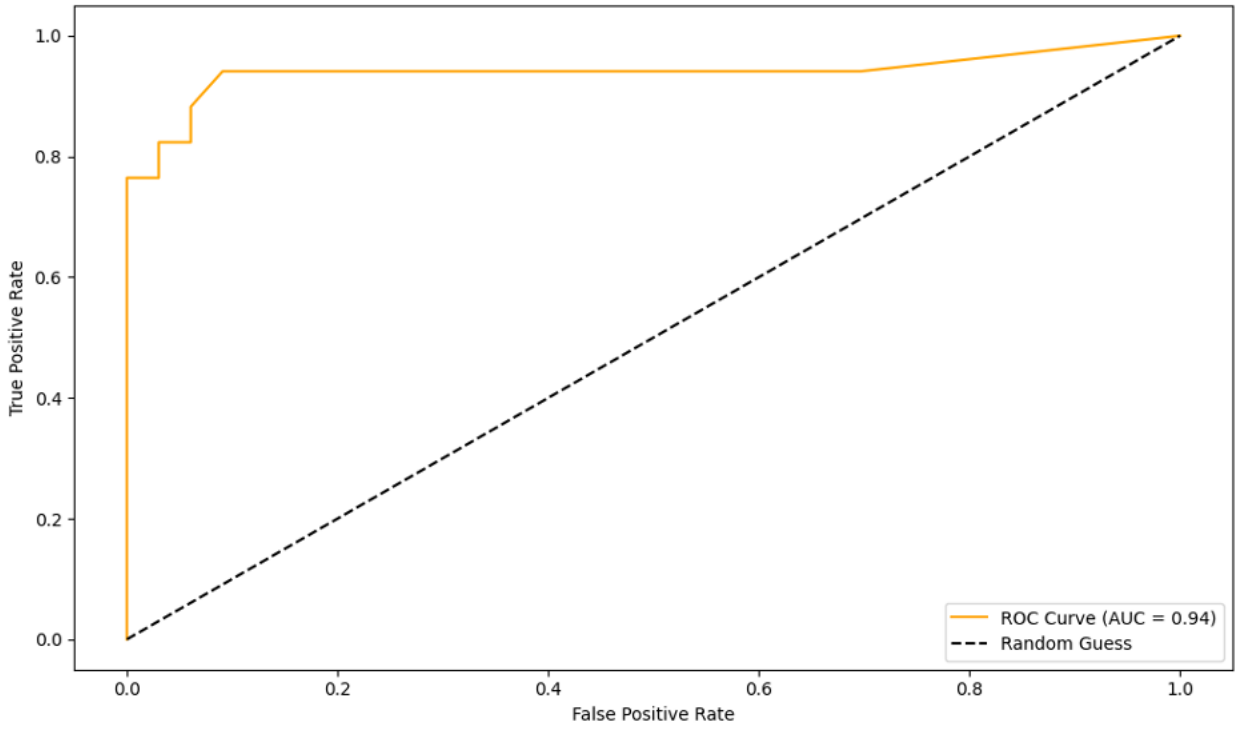

Figure 6 - Receiver Operating Characteristic (ROC) curve for the RF model.

5. Conclusion

The Random Forest classifier demonstrated high efficiency in LULC classification for São Desidério, achieving robust performance, particularly for Cultivated Land, with a test accuracy of 90% and an AUC of 0.94. While slightly reduced recall for Natural Vegetation highlights areas for improvement, such as addressing dataset imbalance or exploring additional spectral indices, the results provide a solid foundation for understanding land use patterns and guiding management strategies. The classification revealed that Cultivated Land accounted for 46.7% of the area, aligning with deforestation patterns and emphasizing São Desidério's role as an agricultural and deforestation hotspot. Future directions include applying Principal Component Analysis (PCA) for dimensionality reduction, testing deep learning, and leveraging time-series analysis to track LULC changes, enabling a deeper understanding of natural and anthropogenic impacts. This scalable, consistent methodology has significant applications in environmental monitoring, policy-making, and sustainable land management.

References

Almeida De Souza, A., Galvão, L. S., Korting, T. S., & Prieto, J. D. (2020). Dynamics of savanna clearing and land degradation in the newest agricultural frontier in Brazil. *GIScience & Remote Sensing*, *57*(7), 965–984. https://doi.org/10.1080/15481603.2020.1835080

Araújo, M. L. S. D., Sano, E. E., Bolfe, É. L., Santos, J. R. N., Dos Santos, J. S., & Silva, F. B. (2019). Spatiotemporal dynamics of soybean crop in the Matopiba region, Brazil (1990–2015). *Land Use Policy*, *80*, 57–67. https://doi.org/10.1016/j.landusepol.2018.09.040

Breiman, L. (2001). Random Forests. *Machine Learning*, *45*(5), 32.

Drusch, M., Del Bello, U., Carlier, S., Colin, O., Fernandez, V., Gascon, F., Hoersch, B., Isola, C., Laberinti, P., Martimort, P., Meygret, A., Spoto, F., Sy, O., Marchese, F., & Bargellini, P. (2012). Sentinel-2: ESA's Optical High-Resolution Mission for GMES Operational Services. *Remote Sensing of Environment*, *120*, 25–36. https://doi.org/10.1016/j.rse.2011.11.026

Ferreira, K. R., Queiroz, G. R., Vinhas, L., Marujo, R. F. B., Simoes, R. E. O., Picoli, M. C. A., Camara, G., Cartaxo, R., Gomes, V. C. F., Santos, L. A., Sanchez, A. H., Arcanjo, J. S., Fronza, J. G., Noronha, C. A., Costa, R. W., Zaglia, M. C., Zioti, F., Korting, T. S., Soares, A. R., … Fonseca, L. M. G. (2020). Earth Observation Data Cubes for Brazil: Requirements, Methodology and Products. *Remote Sensing*, *12*(24), 4033. https://doi.org/10.3390/rs12244033

Fischer, W. A., Hemphill, W. R., & Kover, A. (1976). Progress in remote sensing (1972–1976). *Photogrammetria*, *32*(2), 33–72. https://doi.org/10.1016/0031-8663(76)90013-2

Gislason, P. O., Benediktsson, J. A., & Sveinsson, J. R. (2006). Random Forests for land cover classification. *Pattern Recognition Letters*, *27*(4), 294–300. https://doi.org/10.1016/j.patrec.2005.08.011

Gorelick, N., Hancher, M., Dixon, M., Ilyushchenko, S., Thau, D., & Moore, R. (2017). Google Earth Engine: Planetary-scale geospatial analysis for everyone. *Remote Sensing of Environment*, *202*, 18–27. https://doi.org/10.1016/j.rse.2017.06.031

Organisation for Economic Co-operation and Development, & Food and Agriculture Organization & Food and Agriculture Organization of the United Nations. (2021). *OECD-FAO Agricultural Outlook 2021-2030*. OECD. https://doi.org/10.1787/19428846-en

Russo, G., Alencar, A., Ribeiro, V., Amorim, C., Shimbo, J., Lenti, F., & Castro, I. (2018). *Cerrado: The Brazilian savanna's contribution to GHG emissions and to climate solutions*. IPAM. https://ipam.org.br/wp-content/uploads/2018/12/Policy-Brief-Cerrado-COP24-en-1.pdf

Samuel, A. L. (1959). *Some Studies in Machine Learning Using the Game of Checkers*.

Shao, Y., Lunetta, R. S., Ediriwickrema, J., & Iiames, J. (2010). Mapping Cropland and Major Crop Types across the Great Lakes Basin using MODIS-NDVI Data. *Photogrammetric Engineering & Remote Sensing*, *76*(1), 73–84. https://doi.org/10.14358/PERS.76.1.73

Talukdar, S., Singha, P., Mahato, S., Shahfahad, Pal, S., Liou, Y.-A., & Rahman, A. (2020). Land-Use Land-Cover Classification by Machine Learning Classifiers for Satellite Observations—A Review. *Remote Sensing*, *12*(7), 1135. https://doi.org/10.3390/rs12071135