Enkhjin Munkhbayar

Major: Data Science & Social Systems
Program: Data Science for Social Good (DSSG)
Partner: Stanford Human Trafficking Data Lab

Machine Learning Augmented Prediction for Labor Exploitation Detection in Brazil

Keywords/AMA: machine learning, geospatial analysis, labor exploitation detection, satellite imagery, charcoal production.

Project Overview

Background: Illegal charcoal production in Brazil often involves forced labor. The remote locations and frequent movement of these sites make detection difficult, but satellite imagery offers a promising solution.

Goal: Develop machine learning models to improve the manual identification of suspected charcoal production sites from satellite imagery, enhancing both accuracy and efficiency.

Data: Utilized a geospatial database with high-frequency satellite imagery of Brazil, including confirmed charcoal sites and multi-level spatiotemporal features.

Methods:

🗺️ Identify and integrate new relevant features into the geospatial database.
🔍 Expand labeled training data through detailed inspection and annotation.
🤖 Iteratively train, test, and refine models using geospatial covariates.

Outcome: Developed a model using 38 geospatial features, achieving a 71.7% F1 score, 65.8% precision, and 78.8% recall. Future improvements could include image feature embeddings and enhanced time-series modeling.

Acknowledgements

Thank you to my team members and DSSG Fellows, Leon Reilly and Kyler Shu, our technical mentors, Dr. Benjamin Seiler, Dr. Kimberly Babiarz, and Jonas Junnior, and our faculty mentor, Dr. Michael Baiocchi. We also appreciate the DSSG program for their support and the opportunity to work on this impactful project.

38 Features

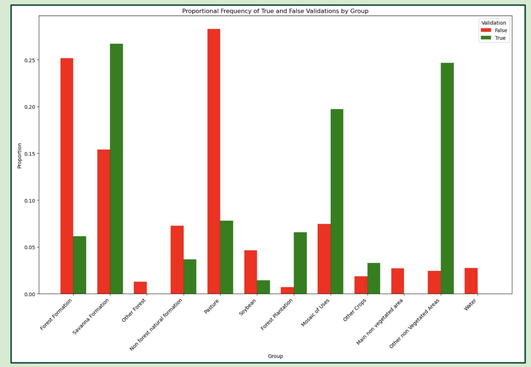
- **14 Distance Variables:** Proximity of sites to key geographical elements
- **3 Density Variables:** Concentration of relevant features around suspected sites
- **12 Land Cover Categories:** Specific land cover group where each site is located
- **9 Survey Variables:** Socioeconomic indicators of regions with suspected sites

Land Cover Feature Insights

Reliable Indicators: Savanna Formation, Other Non-Vegetated Areas, and Mosaic of Uses groups have the highest true positives.

Frequent Misidentifications: Forest Formation and Pasture groups show high false positives.

Unreliable Categories: Other Forest and Water groups have exclusively false positives.



Proportional representation is used to account for the imbalance between false positives and true positives.

Geospatial & Feature Cluster Analysis

Geospatial and feature cluster analyses of charcoal sites were conducted to identify patterns and inform potential improvements in classification methods.