Project Proposal - Group 7

Team:

1. **Tasfia Katha**
2. **Rakshit Mathur**
3. **Nandita Vesangi**
4. **Vishesh Walia**

Executive Summary

This project proposal endeavors to contribute to the global initiative of making science more accessible by addressing the UN Sustainable Development Goal 2: Zero Hunger. By 2022, approximately 735 million people – or 9.2% of the world's population – found themselves in a state of chronic hunger – a staggering rise compared to 2019. This data underscores the severity of the situation, revealing a growing crisis. With around 800 million people facing food insecurity worldwide, exacerbated by factors such as climate change, there is an urgent need to enhance our understanding of crop yields and optimize agricultural practices. Focusing initially on rice production in Vietnam, a country heavily reliant on rice as a staple food source and highly susceptible to the impacts of climate change, this project aims to leverage satellite data to develop a tool for identifying rice crops and use it to work on the goal.

Overview of the problem

As part of a broader initiative to combat food insecurity, this project addresses the pressing challenge of enhancing rice production efficiency, particularly in the face of climate change. With more than 800 million individuals facing hunger globally, and climate change posing significant threats to agricultural systems, there is an urgent need to optimize crop yields. This project focuses on leveraging satellite data to understand and improve rice production, starting with Vietnam due to its status as a major rice producer and its susceptibility to climate-related risks. In Vietnam, where rice is a crucial food source for over half the population, the rice agriculture sector faces challenges like declining yields, the need for higher production due to a growing population, and adverse effects of climate change. Our initiative integrates radar data from Sentinel-1 with high-resolution optical data from Sentinel-2 and Landsat-8,9, using advanced satellite imaging to address these issues.

Approach to solving the problem

Data Acquisition and Preprocessing: Our challenge is to predict the presence of rice crops at 250 geo locations (latitude and longitude) in the An Giang province of Vietnam. We will obtain high-resolution optical data from Sentinel-2 and NASA's Landsat-8,9, along with SAR data from Sentinel-1, covering the Ha Giang province. We will conduct data preprocessing to improve the consistency of the data and model accuracy.

Feature Extraction and Selection: We will extract relevant spectral, textural, and temporal features from preprocessed satellite data. Furthermore, we will combine other bands from the Sentinel dataset to extract more useful information for predicting rice fields. We will try different vegetation indices to understand the presence and health of vegetation such as Normalized Difference Vegetation Index (calculated from the ratio of near-infrared (NIR) and red (R) reflectance), Leaf Area Index (represents the total leaf area per unit ground area for crop yield) etc. Feature selection methods like PCA and RFE will identify discriminative features for rice crop classification.

Model Development and Training: We will use the base logistic model using 600 geolocation training dataset that yielded an F-1 score of 0.55 initially. Afterwards several advanced machine learning algorithms can be applied after conducting exploratory data analysis for instance random forest, support vector machines, artificial neural network etc. We will iterate on each model to see the incremental changes for our model and finalize an algorithm with the highest F-1 score.

Validation and Evaluation: Rigorous validation will be performed using cross-validation techniques and validation datasets. We will perform hyperparameter tuning to optimize the maximum model performance. Finally, metrics like accuracy, precision, recall, and F1-score will be computed to evaluate model reliability.

Resources used to solve the problem

1. A foundational Jupyter notebook, provided by the EY data challenge, serving as an initial template with a preliminary F1 score of 0.55, guiding our exploratory analysis and model development.
2. Using Microsoft's Planetary Computer, we accessed a repository of satellite data and computational resources on the Planetary Hub cloud platform. This enabled seamless data acquisition, processing, and model building within Jupyter Notebooks, facilitating efficient experimentation and iteration.
3. Radar and optical datasets were sourced via Microsoft's Planetary Computer hub, providing a diverse array of data modalities essential for comprehensive analysis.
4. A dataset comprising 250 geolocations of crop fields, enabling targeted analysis and validation of our predictive models against ground truth data.

5. Supplementary notebooks, including tools for Sentinel-2 cloud filtering, Landsat cloud filtering, and Sentinel-1 phenology analysis, augmented our analysis pipeline, enriching our understanding and enhancing the quality of our predictions.
6. EY's comprehensive guide to working with satellite data served as a valuable resource, offering best practices and insights to navigate the complexities of satellite imagery analysis effectively.

Ideas to improve the model's accuracy:

1. **Experiment with More Features:** Apart from VV and VH, we can explore the use of other Sentinel-1 bands or even combine them mathematically to generate new features, such as vegetation indices or texture measures, which can better distinguish between different land classes.
2. **Increase Temporal Resolution:** Expanding the time range from a single day to an entire year can help capture seasonal variability in crop patterns, potentially leading to more accurate predictions.
3. **Feature Aggregation:** Experiment with different spatial aggregation techniques (e.g., bounding boxes) and aggregation functions (e.g., mean, median) to get normalized band values. This can help smooth out noise and improve model performance.
4. **Cross-Validation:** Implement more robust cross-validation techniques, like k-fold cross-validation, to get a more accurate estimate of the model's performance and to reduce overfitting.
5. **Hyperparameter Tuning:** Experiment with tuning the hyperparameters of the models. For example, in logistic regression, we can try different regularization parameters to prevent overfitting and improve model performance.
6. **Ensemble Methods:** Investigate the use of ensemble methods, such as Voting Classifier, Bagging, or Boosting, to combine multiple models and potentially improve the overall performance.
7. **Advanced Algorithms:** Experiment with other machine learning algorithms beyond logistic regression, such as decision trees, random forests, support vector machines, or more advanced deep learning architectures.
8. **Use a More Advanced Model:** We can try more sophisticated models such as XGBoost, or a neural network like a Convolutional Neural Network (CNN) to capture more complex relationships in the data. These models may be better suited to handle the nuances of different land types.

Team member roles for full project

To ensure the success of this project, each team member has been assigned specific roles:

- **Tasfia Katha** and **Rakshit Mathur**: Responsible for both model development and contributing to the value case analysis.
- **Nandita Vesangi**: Tasked with providing regular status updates on project progress.
- **Vishesh Walia**: Engaged in both model development and providing status updates.

Furthermore, the final presentation will be a collective effort, with all team members contributing to effectively communicate the project's outcomes and insights.

References

[1] *"Comprehensive Guide to Satellite Imagery Analysis Using Python." Towards Data Science. Accessed February 10, 2024, from https://towardsdatascience.com/comprehensive-guide-to-satellite-imagery-analysis-using-python-1b4153bad2a.*

[2] Xu et.al., *Paddy Rice Mapping in Thailand Using Time-Series Sentinel-1 Data and Deep Learning Model, Remote Sensing*, 2021.

[3] *"Guidance and Suggestions for Participants of the Data Science Challenge Project." Accessed February 11, 2024, from https://challenge.ey.com/api/v1/storage/admin-files/19107447510343079-65c6679ef59a2bcc0464b217-Comprehensive%20Guide%20v2.docx.*

[4] *United Nations. "Goal 2: Zero Hunger." United Nations Sustainable Development Goals. Accessed February 15, 2024, from https://www.un.org/sustainabledevelopment/hunger/#:~:text=Goal%202%20is%20about%20creating,climate%20change%2C%20and%20deepening%20inequalities.*