Crop Health Classification

Final Report
MALIS - Advanced Project

LABIDI Souha
ZAPATA Juan David

Link to the Github repository

Introduction

Monitoring crop health is crucial for making timely and informed decisions in agriculture. It enables farmers to take quick and effective actions to protect crops, increase yields, and contribute to food security while promoting sustainable farming practices. Traditional methods of crop health assessment, such as field inspections, are often time-consuming, and expensive. These limitations highlight the need for innovative and efficient solutions. Machine learning and remote sensing technologies provide powerful tools to address these challenges. By leveraging real-time satellite imagery, advanced algorithms can analyze crop health more rapidly and accurately than conventional methods. This capability is especially relevant in the context of the Telangana Crop Health Challenge, which focuses on classifying crop types using satellite data. The input to our algorithms includes tabular data enriched with Sentinel-2 satellite imagery, available through the ADeX platform. The tabular data contains essential information on cultivation practices, such as crop type, sowing and irrigation methods, past yields, and crop coverage boundaries. Additionally, the dataset provides .tif format time-series images capturing vegetation indices, soil moisture levels, and crop canopy growth across the growing season, from sowing [1] to harvest [2]. These features are used to predict crop health and inform agricultural decisions. For this project, we implemented and compared the performance of four machine learning algorithms: k-Nearest Neighbors (KNN), Decision Trees, XGBoost, and Logistic Regression. By combining tabular and remote sensing data, we aimed to develop a robust model for predicting crop health and optimizing farming strategies.

[1] Start of season

[2] End of season

Dataset and Features

Dataset Description

The dataset utilized in this project integrates both tabular data and Sentinel-2 satellite imagery to monitor and predict crop health. The tabular data provides information on cultivation practices, including:

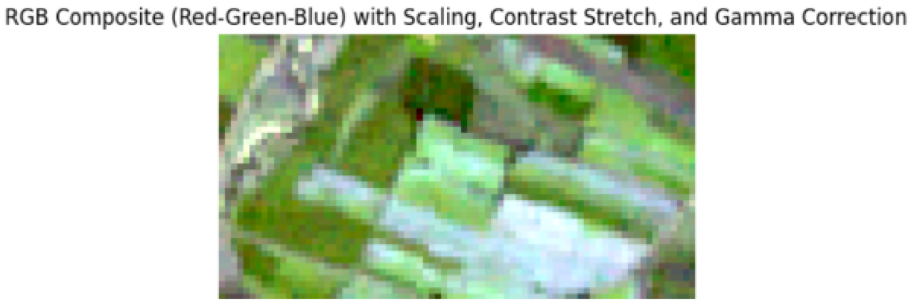
- **Crop type:** Categories of crops being monitored.
- **Sowing and irrigation methods:** Details on planting and watering techniques.
- **Past yield:** Historical data on crop production.
- **Crop coverage:** Defined boundaries of crop fields.

The satellite imagery, in .tif format, captures time-series data over the crop growth cycle, from sowing to harvest. This includes vegetation indices, soil moisture levels, and canopy growth metrics. The dataset is partitioned into:

- **Training set:** 8,775 examples used to train machine learning models.
- **Test set:** 3,016 examples to evaluate model performance.

Data Acquisition

Sentinel-2 images were downloaded, leveraging the ADeX platform. This ensured access to high-resolution, real-time satellite data.



RGB Composite (Red-Green-Blue) with Scaling, Contrast Stretch, and Gamma Correction

Figure 1: Visualization of RGB image from our dataset.

Preprocessing

The following preprocessing steps were applied:

1. **Geometry Encoding:** Encoding spatial information to represent field boundaries and locations effectively.

2. **Feature Engineering:** Deriving relevant features such as:

    - **Data Cleaning and Transformation:** Following the results of the performed exploratory data analysis (EDA), some variables were removed from the dataset due to redundancy or lack of relevance.
    - **Categorical Encoding:** The remaining categorical variables were encoded using a traditional label-encoding approach.
    - **Numerical Feature Standardization:** Numerical variables were standardized using Scikit-learn's `StandardScaler` to ensure uniformity in feature scales.
    - **Date Feature Transformation:** Date-related data was transformed to retain only the month information, reducing the dimensionality while preserving relevant temporal insights.
    - **Vegetation indices** (e.g., NDVI - Normalized Difference Vegetation Index) to assess crop health.

    Through this process the dataset was reduced to only keep 15 attributes.

Image Characteristics

The Sentinel-2 images have a spatial resolution of 10 meters per pixel, suitable for distinguishing detailed crop patterns across fields.

Citation

The dataset was obtained from the ADeX platform, which provides accessible Sentinel-2 imagery and tabular data for agricultural analysis.

Methods

4 different types of algorithms were used to model the data:

1. **K-Nearest Neighbors:** When trying to predict the label of a new point, this algorithm computes the distance to the labeled points (*training data*) to then perform a majority vote between the K closest ones. Due to the curse of dimensionality, only three variables were used (*ndvi,savi and smonth*) and the training set was split into training and validation sets using a 80-20 rule to find the best value for K.

2. **Logistic Regression:** Through a sigmoid function this algorithm computes $P(C \mid x)$ which is the probability of a class happening given a value of x. The parameters for the sigmoid function are found out through a cross-entropy loss function. An under-sampling of the majority class is performed in an attempt to create a model with better generalization capabilities.

3. **Tree Classifier:** This algorithm builds a decision tree by at every node choosing the feature that maximizes an impurity function, that is to say, the feature that separates the data the most. This algorithm is construct with the next three features: *ndvi, savi and water coverage.*

4. **XGBoost:** This boosting technique combines the predictions of weak learners, in this case decision trees, to create a stronger learner. By creating a more complex model, this algorithm is specially suited for those cases where the bias is high and the variance is not a problem. A 80-20 split is carry out to observe how different hyper-parameters combinations affect the model's performance. Different values for the number of estimators and the depth of the trees are studied.

Results

The primary metric used for model evaluation is the weighted average f1-score as it is the one used by the platform and as it is well suited for an unbalanced multi-classification problem like this one. Averaging the f1-scores makes sure the model performs well on every class, and weighting those averages makes sure performing well on the majority class is more important than in the minor classes. The results for the different algorithms are:

1. **K-Nearest Neighbors:** The optimal value of K found is 3 and the f1-score reported by the challenge platform is 0.72533989.

2. **Logistic Regression:** From the models evaluated is this one which presents the worst performance. It does not accomplish the task of learning from data and the f1-score reported by the platform is 0.265174402.

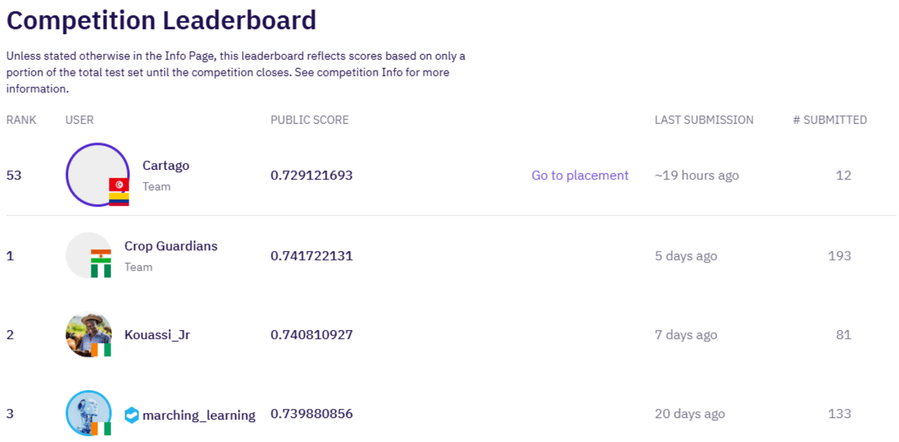
3. **Decision Tree:** This algorithm is not tested in the platform as during its construction it was clear that a stronger learner was required, from the need to use a boosting technique.

4. **XGBoost:** The results obtained do not vary much when using different combinations of hyper-parameters, therefore different submissions with different combinations are done in the platform. The best results are obtained with 100 estimators and a tree depth of 4 with which a f1-score of 0.729121693 is accomplished.

To this day, we rank 53 in the platform from 171 participants. The leader has a f1-score of 0.741722131.

Conclusions

1. When working with an imbalanced dataset, special attention must be put over the metric to be used as simple metrics as the accuracy may not be representing if the model is actually learning from the data and



Competition Leaderboard

Unless stated otherwise in the Info Page, this leaderboard reflects scores based on only a portion of the total test set until the competition closes. See competition Info for more information.

RANK	USER	PUBLIC SCORE		LAST SUBMISSION	# SUBMITTED
53	Cartago Team	0.729121693	Go to placement	~19 hours ago	12
1	Crop Guardians Team	0.741722131		5 days ago	193
2	Kouassi_Jr	0.740810927		7 days ago	81
3	marching_learning	0.739880856		20 days ago	133

Figure 2: Telangana Challenge Leaderboard

understanding its underline distribution. For example, for this project a learner that only assigns the label *Healthy* to every new point could present a very high accuracy despite not really accomplishing the task of detecting unhealthy crops.

2. The combination of the remote sensing and machine learning technologies can create a powerful tool for real-time prediction of natural phenomena. However, the construction of a well built ground truth data remains an expensive task.

3. Despite being usually underestimated, the K-Nearest Neighbors algorithm remains a powerful technique for classification tasks in low dimensions. Its intuition based on looking for the lowest distances makes it suitable even for multiclasses problems.

4. Boosting techniques help increasing the complexity of the built learner which can be useful to solve under-fitting situations.

5. Observing a learner that performs much better in the training set than in the test one is a clear sign of over-fitting. For this cases, reducing the learner complexity is one of the best ways to go in order to find a model with better generalization capabilities. This was clear when XGBoost models created with decision trees of depth 4 clearly outperformed trees of depth 5 and 6 when competing in the challenge.

Contributions

Souha did the satellite data collection while Juan the univariate and bivariate analysis. We both worked together for the feature engineering and model training steps.